

# DBSCAN: Optimal Rates For Density Based Clustering

Daren Wang, Xinyang Lu, Alessandro Rinaldo

## Abstract

We study the problem of optimal estimation of the density cluster tree under various assumptions on the underlying density. Building up from the seminal work of [Chaudhuri et al. \[2014\]](#), we formulate a new notion of clustering consistency which is better suited to smooth densities, and derive minimax rates of consistency for cluster tree estimation for Hölder smooth densities of arbitrary degree  $\alpha$ . We present a computationally efficient, rate optimal cluster tree estimator based on a straightforward extension of the popular density-based clustering algorithm DBSCAN by [Ester et al. \[1996\]](#). The procedure relies on a kernel density estimator with an appropriate choice of the kernel and bandwidth to produce a sequence of nested random geometric graphs whose connected components form a hierarchy of clusters. The resulting optimal rates for cluster tree estimation depend on the degree of smoothness of the underlying density and, interestingly, match minimax rates for density estimation under the supremum norm. Our results complement and extend the analysis of the DBSCAN algorithm in [Sriperumbudur and Steinwart \[2012\]](#). Finally, we consider level set estimation and cluster consistency for densities with jump discontinuities, where the sizes of the jumps and the distance among clusters are allowed to vanish as the sample size increases. We demonstrate that our DBSCAN-based algorithm remains minimax rate optimal in this setting as well.

## 1 Introduction

Clustering is one of the most basic and fundamental tasks in statistics and machine learning, used ubiquitously and extensively in the exploration and analysis of data. The literature on this topic is vast, and practitioners have at their disposal a multitude of algorithms and heuristics to perform clustering on data of virtually all types. However, despite its importance and popularity, rigorous statistical theories for clustering, leading to inferential procedures with provable theoretical guarantees, have been traditionally lacking in the literature. As a result, the practice of clustering, one of the most central tasks in the analysis and manipulation of data, still relies in many cases on methods and heuristics of unknown or even dubious scientific validity. One of the most striking instances of such a disconnect is the algorithm DBSCAN of [Ester et al. \[1996\]](#), an extremely popular and relatively efficient [see [Gan and Tao, 2015](#), [Wang et al., 2015](#)] clustering methodology whose statistical properties have been properly analyzed only very recently in [Sriperumbudur and Steinwart \[2012\]](#).

In this paper, we provide a complementary and thorough study of DBSCAN, and show that this simple algorithm can deliver optimal statistical performance in *density-based* clustering. Density-based clustering [see, e.g., [Hartigan, 1981](#)] provides a general and rigorous probabilistic framework in which the clustering task is well-defined and amenable to statistical analysis. Given a Borel probability distribution  $P$  on  $\mathbb{R}^d$  with Lebesgue density  $p$  and a fixed threshold  $\lambda \geq 0$ , the  $\lambda$ -clusters of  $p$  are the connected components of the upper  $\lambda$ -level set of  $p$ , the set  $\{x \in \mathbb{R}^d: p(x) \geq \lambda\}$

of all points whose density values exceed the level  $\lambda$ . With this definition, clusters are the high-density regions, subsets of the support of  $P$  with the largest probability content among all sets of the same volume.

As noted in [Hartigan \[1981\]](#), the hierarchy of inclusions of all clusters of  $p$  is a tree structure indexed by a height or level parameter  $\lambda > 0$ , called the *cluster tree* of  $p$ . The chief goal of density clustering is to estimate, given an i.i.d. sequence  $\{X_i\}_{i=1}^n$  of points from  $P$ , the cluster tree of  $p$ . A density tree clustering estimator is also a tree structure, consisting of a hierarchy of nested subsets of the sample points, and typically relies on non-parametric estimators of  $p$  in order to determine which sample points belong to high-density regions of  $p$ . A cluster tree estimator is deemed accurate if, with high probability, the hierarchy of clusters it encodes is close, in an appropriate sense, to the hierarchy that would have been obtained should  $p$  be known.

Density-based clustering is an instance of hierarchical clustering that enjoys several advantages: (1) it imposes virtually no restrictions on the shape, size and number of clusters, at any level of tree; (2) unlike *flat* (i.e. non-hierarchical) clustering, it does not require as input a pre-specified number of clusters and in fact the number of clusters itself is quantity that may change depending on the level of the tree; (3) it provides a multi-resolution representation of all the clustering features of  $p$  across all levels  $\lambda$  at the same time; (4) it allows for an efficient encoding of the entire tree of clusters with a compact data structure that can be easily accessed and queried, and (5) the main object of interest for inference, namely the cluster tree of  $p$ , is a well-defined population quantity, and the notions of consistency of a cluster tree estimator and of its uncertainty are well-defined.

Despite the appealing properties of the density-based clustering framework, a rigorous quantification of the statistical performance of such algorithms has proved difficult. The seminal work by [Hartigan \[1981\]](#) and then by [Penrose \[1995\]](#) has led to a relatively weak notion of cluster consistency, achieved by the popular single-linkage algorithm. More recently [Chaudhuri et al. \[2014\]](#) have developed a general framework for defining consistency of cluster tree estimators based on a separation criterion among clusters. The authors have further demonstrated two algorithms, both based on  $k$ -nearest neighbors graphs over the sample points, that achieved such consistency. One of these algorithms was also shown to achieve the optimal minimax scaling with respect to certain parameters that quantify the hardness of the clustering task. Those results have been generalized in [Balakrishnan et al. \[2012\]](#), where it is shown that the main algorithm of [Chaudhuri et al. \[2014\]](#), as well a class of kernel density estimators, ensure similar consistency guarantees for cluster trees arising from probability distributions supported over well-behaved manifolds, with consistency rates depending on the reach of the manifold and its intrinsic dimension. In both contributions, rates for cluster consistency are established with virtually no assumptions on the underlying density. In particular, these rates do not directly reflect the degree of smoothness of the underlying density.

Below, we will provide further contributions to the theory of density based clustering by deriving new and minimax optimal rates of consistency for cluster tree estimation that depend explicitly on the smoothness of the underlying density function. In line with well-known results from the minimax theory of density estimation, we establish that the cluster trees of smoother densities can be consistently estimated at faster rates that depend on the smoothness of the density. Interestingly, such rates match those for estimating smooth densities in the  $L_\infty$  norm. To the best of our knowledge, this finding and the implication that density based clustering is no easier – at least in our setting – than density estimation, has not been rigorously shown before. In order to account explicitly for the smoothness of the density, we have developed a new criterion for cluster consistency that is better suited for smooth densities. In terms of procedures, we consider cluster

tree estimators that arise from applying a very simple generalization to the well-known DBSCAN procedure and are computationally efficient. Despite its simplicity, our DBSCAN-based estimator is minimax optimal over arbitrary smooth densities under our notion of consistency and under appropriate conditions.

Overall our contributions further advance our theoretical understanding of density-based clustering.

## Problem-Set-up

Let  $P$  be a Borel probability distribution supported on  $\Omega \subset \mathbb{R}^d$  and with Lebesgue density  $p$ . Notice that, necessarily,  $\Omega$  has dimension  $d$ .

**Definition 1.** For any  $\lambda \geq 0$ , let  $L(\lambda) = \{x \in \Omega: p(x) \geq \lambda\}$  be the  $\lambda$ -upper level set of  $p$ . For a given  $\lambda \geq 0$ , the  $\lambda$ -cluster of  $p$  are the connected components of  $L(\lambda)$ .

We refer the reader to Appendix A for a definition of connectedness. Notice that the set of all clusters is an indexed collection of subsets of  $\Omega$ , whereby each cluster of  $p$  is assigned the index  $\lambda$  associated to the corresponding super-level set  $L(\lambda)$ , and that many clusters may be indexed by the same level  $\lambda$ .

**Definition 2.** The cluster tree of  $p$  is the collection  $T_p$  of all clusters of  $p$ , indexed by  $\lambda \geq 0$ . We can represent the cluster tree of  $p$  as the function on  $[0, \infty)$  returning, for each  $\lambda \geq 0$ , the set of  $\lambda$ -clusters of  $p$ .

Thus,  $T_p(\lambda)$  consists of disjoint connected subsets of  $\Omega$  or is empty. In particular,  $T_p(0) = \Omega$  and  $T_p(\lambda) = \emptyset$  if and only if  $\lambda > \|p\|_\infty := \sup_{x \in \Omega} p(x)$ .

The cluster tree owes its name to the easily verifiable property [see Hartigan, 1981] that if  $A$  and  $B$  are elements of  $T_p$ , i.e. distinct clusters of  $p$ , then  $A \cap B = \emptyset$  or  $A \subseteq B$  or  $B \subseteq A$ . This induces a partial order on the set of clusters. In particular, for any  $\lambda_1 \geq \lambda_2 \geq 0$ , if  $A \in T_p(\lambda_1)$  and  $B \in T_p(\lambda_2)$  then either  $A \cap B = \emptyset$  or  $B \subseteq A$ . As a result,  $T_p$  can be represented as a dendrogram with height indexed by  $\lambda \geq 0$ . We refer the reader to Kim et al. [2016] for a formal definition of the dendrogram encoding a cluster tree.

Let  $\{X_i\}_{i=1}^n$  be an i.i.d. sample from  $P$ . In order to estimate the cluster tree of  $p$  we will consider tree-valued estimators, defined below.

**Definition 3.** A cluster tree estimator of  $T_p$  is a collections  $\widehat{T}_n$  of subsets of  $\{X_i\}_{i=1}^n$  indexed by  $[0, \infty)$  such that

- for each  $\lambda \geq 0$ ,  $\widehat{T}_n(\lambda)$  is either empty or consists of disjoint subsets of  $\{X_i\}_{i=1}^n$ , called clusters, and
- $\widehat{T}_n$  satisfies the following tree property: for any  $\lambda_1 \geq \lambda_2 \geq 0$ , if  $A \in \widehat{T}_n(\lambda_1)$  and  $B \in \widehat{T}_n(\lambda_2)$  then either  $A \cap B = \emptyset$  or  $B \subseteq A$ .

It is important to realize that while the cluster tree  $T_p$  of density  $p$  is a collection of connected subsets of its support, the cluster tree estimators considered in this paper are comprised by collections of subsets of the sample points partially ordered with respect to the inclusion relation.

In order to quantify how well a cluster-tree estimator approximates the true cluster tree, we will rely on the notion of cluster tree consistency put forward by Chaudhuri et al. [2014], which we rephrase next.

**Definition 4.** Given data  $\{X_i\}_{i=1}^n$ , let  $\{\mathcal{A}_n\}_{n=1}^\infty$  denote a sequence of collections of connected subsets of the support of  $p$ . A cluster tree estimator  $\widehat{T}_n$  is consistent with respect to the sequence  $\{\mathcal{A}_n\}_{n=1}^\infty$  if, as  $n$  tends to infinity, the following holds, simultaneously over all disjoint elements  $A$  and  $A'$  in  $\mathcal{A}_n$ : with probability tending to 1, the smallest clusters in  $\widehat{T}_n$  containing  $A \cap \{X_i\}_{i=1}^n$  and  $A' \cap \{X_i\}_{i=1}^n$  are disjoint.

In this definition, the collection  $\mathcal{A}_n$  may include clusters, but also other connected subsets of the support of  $p$ . The requirement for consistency outlined above is rather natural: if a cluster tree is to be deemed consistent with respect to the sequence  $\mathcal{A}_n$ , then it should, with probability tending to 1, cluster the sample points perfectly well. Or equivalently, as well as if we had ability of verifying, for each pair of sample points  $X_i$  and  $X_j$  and each connected set  $A \in \mathcal{A}_n$ , whether both  $X_i$  and  $X_j$  are in  $A$ . We take notice that the above definition only requires  $\widehat{T}_n$  to preserve the connectivity of all the sets in  $\mathcal{A}_n$ . However,  $\widehat{T}_n$  might have additional unwanted clusters, referred to as *spurious* in Chaudhuri et al. [2014] that do not correspond to any disjoint sets in  $\mathcal{A}_n$ . We will come back to this in Section 3.7.

The reason why we consider a sequence  $\{\mathcal{A}_n\}_{n=1}^\infty$  of collections of connected sets is to allow the set of target connected subsets of  $p$ , such clusters, to grow larger and more complex as the sample size  $n$  increases, so that the cluster tree estimator will be able, as more data are collected, to discriminate among clusters of  $p$  that are barely distinguishable. An example of a sequence  $\{\mathcal{A}_n\}$  is the set of  $\delta_n$ -separated clusters according to Definition 5 below, where the parameter  $\delta_n$  is taken to be vanishing in  $n$ .

The sequence of target subsets  $\{\mathcal{A}_n\}_{n=1}^\infty$  may not be chosen to be too large: for example if  $\mathcal{A}_n$  is equal, for each  $n$ , to the set of all clusters of  $p$ , then, depending on the complexity of  $p$ , no cluster tree estimator need to be consistent. A natural way to define  $\{\mathcal{A}_n\}_{n=1}^\infty$  is by specifying a *separation criterion* for sets, which may become less strict as  $n$  grows, and then populate each  $\mathcal{A}_n$  using only the connected subsets of the support of  $p$  fulfilling such a criterion. In particular, Chaudhuri et al. [2014] develop a separation criteria known as the  $(\epsilon, \sigma)$ -separation, which requires two subsets connected  $A$  and  $A'$  to be far apart from each other in terms of their “horizontal” distance  $d(A, B) = \inf_{x \in A, y \in B} \|x - y\|$  and their “vertical” distance, in the sense that the smallest cluster containing both  $A$  and  $B$  should belong to a level set of  $p$  indexed by a value of  $\lambda$  significantly smaller by the values indexing the level sets of  $A$  and  $B$ . See Definition 11 below for details. One of the major contributions in this paper is to replace this rather general notion of separation, which requires the specification of two independent parameters, by a simpler one – the  $\delta$ -separation criterion of Definition 5 – which is natural for smooth densities and allows to extract faster cluster rates.

## Related Work

The idea of using probability density function to study clustering structure dates back to Hartigan [1981], who formalized the notion of clusters as the connected components of the high density regions. This formalism was later explored by many. Among others, Rinaldo and Wasserman [2010], Polonik [1995] focus on the clustering consistency of a fixed level; Stuetzle and Nugent [2010], Stuetzle [2003] analyze efficient tree algorithms; Rinaldo et al. [2012] investigate the stability of the clustering structure; Eldridge et al. [2015], Kim et al. [2016] study the inference of the trees under various tree metric. Recently, Chaudhuri et al. [2014] proposes a simple algorithm and show that with appropriate choice of the parameters, the resulting hier-

archical clustering structure correctly estimates the cluster tree with high probability. Based on their results, [Kpotufe and Luxburg \[2011\]](#) further proposed efficient pruning algorithms. (See also [Klemelä \[2009\]](#).)

The density level sets estimation and support estimation have also been intensively studied in the statistic literature. A comprehensive summary of the early works on the support estimation can be found in [Tsybakov et al. \[1997\]](#). Different approaches are later studied by many authors (see e.g, [Ba et al. \[2000\]](#), [Cuevas and Fraiman \[1997\]](#), [Klemelä \[2004\]](#)). Being a closely related topic, the level set estimation received a lot more attention in the recent years. For example, [Cuevas et al. \[2006\]](#), [Tsybakov et al. \[1997\]](#) focus on the consistency, [Rigollet and Vert \[2009\]](#), [Willett and Nowak \[2007\]](#) analyze the minimaxity under various loss functions, [Chen et al. \[2016\]](#) discusses inference and visualization and [Singh et al. \[2009\]](#) investigates the adaptive histogram estimator, and show its optimality. [Jiang \[2017b\]](#) studies the uniform convergence rates for kernel density estimator and [Jiang \[2017a\]](#) analyzes density level set estimation on manifolds using DBSCAN.

## Summary of Our Contributions

We briefly summarize of the main contribution of our manuscript.

- In Section 3 we study cluster tree density estimation of Hölder continuous densities or arbitrary smoothness  $\alpha > 0$ . We formulate a novel criterion of separation among connected subsets that lead to a new notion of cluster consistency, called  $\delta$ -consistency. We exhibit cluster tree estimators in Algorithm 1 (for the case of  $\alpha \leq 1$ ) and Algorithm 2 (for the case of  $\alpha > 1$ ) that are computationally efficient and minimax optimal for cluster tree estimation. We show that the optimal rates of cluster consistency depend on the degree of smoothness of the underlying density and are, up to logarithmic factors, the same rates for estimating an Hölder-smooth density in the sup-norm loss. This result implies that, for the class of densities under consideration, clustering is as difficult as density estimation in the sup-norm. Though not surprisingly, this result has not been previously established.
- In Section 4 we consider the different scenario in which the underlying density exhibits jump discontinuities. We are particular interested in clustering consistency right below the density level at which the jump occurs, and assuming that the size of the discontinuity is vanishing in  $n$  (so that clustering becomes increasingly difficult). We show that with suitable inputs, the DBSCAN algorithm returns a Devroye-Wise type of estimator which is minimax optimal for cluster recovery and level set estimation. The main contribution on this section is to derive the minimax scaling for the size of the jump discontinuity, which appears to not have been previously known.

Our analysis is based on finite sample bounds. We have made an attempt to keep track of the constants in most of the bounds and of their dependence on other fixed quantities such as the dimension and other properties of the underlying density. While it would be desirable to allow for a dimension changing with  $n$ , this modification will add significant complexity to the problem and will require a separate analysis. Thus we have followed the convention commonly adopted in the literature on density cluster and density estimation and have treated  $d$  as fixed.

## Notation

Throughout, we denote with  $p$  the underlying Lebesgue density of the i.i.d. sample  $\{X_i\}_{i=1}^n \subset \mathbb{R}^d$ . We let  $\mathcal{L}_d$  be the Lebesgue measure in  $\mathbb{R}^d$  and, for a point  $x \in \mathbb{R}^d$  and a value  $r > 0$ , we let  $B(x, r)$  be the  $d$ -dimensional closed Euclidean ball centered at  $x$  and with radius  $r$ . We write  $V_d = \mathcal{L}_d(B(0, 1))$  for the volume of the Euclidean unit ball  $B(0, 1)$ . For a real valued function on  $\mathbb{R}^d$  we set  $\|f\|_\infty = \sup_x |f(x)|$  for its  $L_\infty$  (supremum) norm. For a Lebesgue density  $p$  on  $\mathbb{R}^d$  and  $\lambda \geq 0$ , we let  $L(\lambda) = \{x \in \mathbb{R}^d : p(x) \geq \lambda\}$  stand for its upper level set at  $\lambda$ . We use  $T_p$  to denote the cluster tree with underlying density  $p$  and  $\widehat{T}_n$  to denote any estimator of the cluster tree (see Definitions 2 and 3 above). For any measurable set  $A \subset \mathbb{R}^d$  and any  $h > 0$ , we define

$$A_h = \bigcup_{x \in A} B(x, h) \quad \text{and} \quad A_{-h} = \{x \in A : B(x, h) \subset A\}. \quad (1)$$

Throughout, we will denote with  $C, C_1, C_2$  quantities that do not depend on any variable of interest and whose value may change from line to line. These constants may depend on other parameters held fixed, such as the dimension  $d$ . We will indicate such dependence but not track it explicitly in our statements.

## 2 The DBSCAN Algorithm and its Connections with KDE

In this section we describe how straightforward generalization of the DBSCAN algorithm of Ester et al. [1996] will produce a cluster tree estimator and elucidate its connections with kernel density estimation; see Algorithm 1. As shown below in Section 3.4.1, the resulting estimator will be optimal provided that the underlying density is Hölder continuous with parameter  $\alpha \leq 1$  (see Section 3.1 for a definition of Hölder continuity). For higher degrees of smoothness, it will be necessary to utilize a slight variant of this procedure, given in Algorithm 2, in order to retain optimality.

---

**Algorithm 1** The DBSCAN algorithm

---

**INPUT:** i.i.d sample  $\{X_i\}_{i=1}^n$  and  $h > 0$ . For each  $k \in \{0, \dots, n\}$ ,

1. construct a graph  $\mathbb{G}_{h,k}$  with node set  $\{X_i : |B(X_i, h) \cap \{X_j\}_{j=1}^n| \geq k\}$  and edge set  $\{(X_i, X_j) : \|X_i - X_j\| < 2h\}$ ;
3. compute  $\mathbb{C}(h, k)$ , the maximal connected components of  $\mathbb{G}_{h,k}$ .

**OUTPUT:**  $\widehat{T}_n = \{\mathbb{C}(h, k), k \in \{0, \dots, n\}\}$

---

It is easy to see that the output of Algorithm 1 is a cluster tree estimator according to Definition 3. Indeed, for a given value of  $k \in \{0, \dots, n\}$  and  $h > 0$ , the subset  $\mathbb{C}(h, k)$  of sample points (if non-empty) correspond to a “flat” clustering. By sweeping through all the possible values of  $k$  the algorithm returns a sequence of nested geometric graphs over the sample points. The hierarchy of connected components of such graphs is then a cluster tree estimator since, for each pair of integers  $k_1 \leq k_2$ ,

$$\bigcup_{\{X_i : |B(X_i, h) \cap \{X_j\}_{j=1}^n| \geq k_2\}} B(X_i, h) \subseteq \bigcup_{\{X_i : |B(X_i, h) \cap \{X_j\}_{j=1}^n| \geq k_1\}} B(X_i, h).$$

**Remark 1.** In practice, Algorithm 1 can be efficiently implemented using a union-find structure in such a way that the determination of the set  $\mathbb{C}(h, k)$  of maximal connected components of  $\mathbb{G}_{h,k}$

can be accomplished without using the potentially expensive breadth-first search or depth-first search algorithms. See [Najman and Couprie \[2006\]](#) for an efficient implementation.

For a fixed value of  $k$ , Algorithm 1 is in fact a slightly simplified version of in the original DBSCAN algorithm of [Ester et al. \[1996\]](#) for “flat” clustering”, where the parameters  $h$  and  $k$  are called instead Eps and MinPts, respectively. Specifically, in the original version of DBSCAN, two nodes  $X_i$  and  $X_j$  in the graph  $\mathbb{G}_{h,k}$  are connected if  $\|X_i - X_j\| < h$  instead of  $2h$ . Such a variant allows us to link the connected components of the graph  $\mathbb{G}_{h,k}$  to the connected components of the upper level set  $\{\hat{p}_h \geq \lambda_k\}$  of the density estimator  $\hat{p}_h$  (see Equation (2) below), where  $\lambda_k = \frac{k}{nh^d V_d}$ . This simplifies our theoretical analysis without affecting the consistency rates. A second minor difference is the fact that Algorithm 1 does not distinguish between *core* and *border* points. It is possible to show that such a distinction is also largely inconsequential in deriving consistency rates for Algorithm 1 and, therefore, we have not included it in our analyst.

As pointed out by [Sriperumbudur and Steinwart \[2012\]](#), DBSCAN corresponds to using a kernel density estimator with a spherical kernel  $K$  given by the indicator function of the unit  $d$ -dimensional Euclidean ball to cluster the points. In detail, consider the density estimator  $\hat{p}_h: \mathbb{R}^d \rightarrow \mathbb{R}$  given by

$$x \mapsto \hat{p}_h(x) = \frac{|B(x, h) \cap \{X_i\}_{i=1}^n|}{nh^d V_d} = \frac{1}{nh^d V_d} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right), \quad (2)$$

where

$$K(x) = \begin{cases} 1 & \text{if } x \in B(0, 1) \\ 0 & \text{otherwise.} \end{cases} \quad x \in \mathbb{R}^d. \quad (3)$$

It is easy to see that  $\hat{p}_h$  is a Lebesgue density, i.e.  $\hat{p}_h \geq 0$  for all  $x$  and  $\int_{\mathbb{R}^d} \hat{p}_h(x) dx = 1$ . Furthermore,

$$\mathbb{E}[\hat{p}_h(x)] = p_h(x), \quad \forall x \in \mathbb{R}^d, \quad (4)$$

where

$$p_h(x) = \frac{1}{h^d V_d} \int_{\mathbb{R}^d} K\left(\frac{x - z}{h}\right) p(z) dz = \frac{P(B(x, h))}{h^d V_d}. \quad (5)$$

The connections between DBSCAN and the density estimator  $\hat{p}_h$  are described in the following result, whose proof is immediate. For any  $\lambda \geq 0$ , set

$$\hat{D}(\lambda) = \{x: \hat{p}_h(x) \geq \lambda\} \cap \{X_i\}_{i=1}^n$$

and

$$\hat{L}(\lambda) = \bigcup_{X_j \in \hat{D}(\lambda)} B(X_j, h) \quad (6)$$

**Lemma 1.** *Let  $k$  and  $h$  be the input to DBSCAN. Then the nodes of  $\mathbb{G}_{h,k}$  is the set  $\hat{D}(\lambda_k)$  where  $\lambda_k = \frac{k}{nh^d V_d}$ . Furthermore, two points  $X_i$  and  $X_j$  in  $\hat{D}(\lambda_k)$  are in the same connected component of  $\hat{L}(\lambda_k)$  if and only if they are in the same graphical connected component of  $\mathbb{G}_{h,k}$ . Consequently, for any pair  $A$  and  $A'$  of subsets of  $\mathbb{R}^d$  with  $A \cap \{X_i\}_{i=1}^n \neq \emptyset$  and  $A' \cap \{X_i\}_{i=1}^n \neq \emptyset$ ,*

- if  $A \subset \hat{L}(\lambda_k)$  is connected, all the sample points in  $A$  belong to the same connected component of  $\mathbb{G}_{h,k}$ .

- if  $A$  and  $A'$  belongs to distinct connected components of  $\widehat{L}(\lambda_k)$ , then the sample points in  $A$  and the sample points in  $A'$  belong to distinct connected components of  $\mathbb{G}_{h,k}$ .

Notice that while  $\widehat{D}(\lambda)$  is a finite collection of points in  $\mathbb{R}^d$ ,  $\widehat{L}(\lambda)$  is a  $d$ -dimensional closed set. For clustering purposes however, the two sets convey the same information. From a computational standpoint, this equivalence is key to the efficiency of Algorithm 1: it is computationally very inexpensive to check whether two points  $X_i$  and  $X_j$  are in the same connected component of the graph  $\mathbb{G}(h, k)$  (something that, as remarked above, follows directly from a simple union-find strategy); in contrast, checking whether  $X_i$  and  $X_j$  are in the same connected component of  $\{x \in \mathbb{R}^d : \widehat{p}_h(x) \geq \lambda\}$  can be computationally costly, even for small values of  $d$ .

The estimator  $\widehat{L}(\lambda)$  is well known in the literature on level set estimation; e.g., Cuevas and Rodríguez-Casal [2004] and Devroye and Wise [1980]. Furthermore, as shown in Sriperumbudur and Steinwart [2012], with a suitable scaling of the bandwidth parameter  $h$  and under appropriate assumptions on the underlying density  $p$  and/or its support,  $\widehat{L}(\lambda)$  is a rate optimal minimax estimator of the level set  $\{p \geq \lambda\}$ .

### 3 Clustering rate for Hölder continuous densities

In this section we study optional estimation of the density cluster tree when the density is Hölder smooth. In this case, the notion of  $(\epsilon, \sigma)$ -separation originally proposed by Chaudhuri et al. [2014] to quantify the discrepancy between clusters can be refined significantly, since the smoothness properties of the underlying density constraint the range of the possible combinations of the vertical and horizontal separation  $\epsilon$  and  $\sigma$  between clusters. In fact, we will show below in Section 3.2 that a simpler notion of cluster separation, called  $\delta$ -separation, which depends only on one parameter, will suffice to capture the inherent difficulty of density clustering under smoothness of the density. We will comment on the differences between the  $(\epsilon, \sigma)$ -separation and the  $\delta$ -separation criteria below in Section 5.1.

#### 3.1 Hölder Smooth Densities

Given vectors  $s = (s_1, \dots, s_d)$  in  $\mathbb{N}^d$  and  $x = (x_1, \dots, x_d)$  in  $\mathbb{R}^d$ , set  $|s| = s_1 + \dots + s_d$  and  $x^s = x_1^{s_1} \dots x_d^{s_d}$ , and let

$$D^s = \frac{\partial^{s_1 + \dots + s_d}}{\partial x_1^{s_1} \dots \partial x_d^{s_d}}$$

denote the high-order differential operator. A Lebesgue density  $p : \mathbb{R}^d \rightarrow \mathbb{R}$  is said to belong to the Hölder class  $\Sigma(L, \alpha)$  with parameters  $\alpha > 0$  and  $L > 0$  if  $p$  is  $\lfloor \alpha \rfloor$ -times continuously differentiable and, for all  $x, y \in \mathbb{R}^d$  and all  $s \in \mathbb{N}^d$  with  $|s| = \lfloor \alpha \rfloor$ ,

$$|D^s p(x) - D^s p(y)| \leq L \|x - y\|^{\alpha - |s|}.$$

Notice that, when  $0 < \alpha \leq 1$ , the Hölder condition reduces to the Lipschitz condition

$$|p(x) - p(y)| \leq L \|x - y\|^\alpha, \quad \forall x, y \in \mathbb{R}^d.$$

In our analysis below, we will require a high-probability bound on the quantity

$$\|\widehat{p}_h - p\|_\infty \leq \|\widehat{p}_h - p_h\|_\infty + \|p_h - p\|_\infty$$



where  $p$  is assumed to belong to the class  $\Sigma(L, \alpha)$ , for some  $L > 0$  and  $\alpha > 0$ , and  $\hat{p}_h$  is a kernel density estimator with bandwidth  $h > 0$  of the form

$$x \mapsto \frac{1}{nh^d} \sum_{i=1}^d K\left(\frac{x - X_i}{h}\right)$$

for some kernel  $K$ , such as the one given Equation (3) for example, and  $p_h(x) = \mathbb{E}[\hat{p}_h(x)]$ , for all  $x \in \mathbb{R}^d$ . As is customary in non-parametric density estimation, we obtain separate bounds for the stochastic component  $\|\hat{p}_h - p_h\|_\infty$  and the bias  $\|p_h - p\|_\infty$ . For the first term we will invoke known concentration bounds from density estimation to conclude that there exists a quantity  $C_1$  such that, for any  $\gamma > 0$  and assuming  $nh^d \geq 1$ ,

$$\mathbb{P}\left(\|\hat{p}_h - p_h\|_\infty \leq \frac{C_1(\gamma + \log(1/h))}{\sqrt{nh^d}}\right) \geq 1 - e^{-\gamma}. \quad (7)$$

The verification of the previous bound is given in Appendix B. There, we distinguish two cases. If  $\alpha \leq 1$  we may take the kernel  $K$  to be the spherical kernel as in Equation (3), so that  $\hat{p}_h$  will reduce to Equation (2) and we are effectively recovering the DBSCAN procedure. In this case, the constant  $C_1$  will only depend on  $\|p\|_\infty$  and  $d$ ; see Proposition 16 in Appendix B.

When instead  $\alpha > 1$ , Equation (7) holds provided that the kernel  $K$  satisfies the so-called VC property, which we recall in Appendix B. The VC-property is verified for a large class of kernels, including any compact supported polynomial kernel and the Gaussian kernel. See Nolan and Pollard [1987] and Giné and Guillou [2002]. In this case, the constant  $C_1$  will additionally depend on the VC characteristic of  $K$  (see Proposition 17 in Appendix B for details). As for the bias term  $\|p_h - p\|_\infty$ , standard calculations yield that, for an appropriate constant  $C_2$ ,

$$\|p_h - p\|_\infty \leq C_2 h^\alpha. \quad (8)$$

When  $\alpha \leq 1$ ,  $C_2$  depends on  $L$  only. When  $\alpha > 1$ , Equation (8) holds for a certain class of kernels known as  $[\alpha]$ -valid kernels, whose construction can be found, e.g., in Rigollet and Vert [2009]. Since this type of kernels are polynomials supported on  $[0, 1]^d$ , they automatically satisfy the VC condition. See lemma 22 of Nolan and Pollard [1987] for instance. We remark that for  $\alpha > 2$ ,  $[\alpha]$ -valid kernels take on negative values. In this case  $C_2$  depends on  $L$ ,  $K$  and  $\alpha$ .

Thus combining the bias and the variance bounds (7) and (8), we conclude that, for any  $\gamma > 0$ , with probability at least  $1 - e^{-\gamma}$ ,

$$\|\hat{p}_h - p\|_\infty \leq a_n, \quad (9)$$

where  $a_n = \frac{C_1(\gamma + \log(1/h))}{\sqrt{nh^d}} + C_2 h^\alpha$  and the kernel  $K$  may be taken to be the spherical kernel if  $\alpha \leq 1$  and is an  $[\alpha]$ -valid, VC kernel otherwise. Setting  $\gamma = \log n$ , the optimal choice of the bandwidth is

$$h \asymp \frac{\log n}{n^{2\alpha+d}}, \quad (10)$$

which leads the rate

$$\|\hat{p}_h - p\|_\infty \leq C \left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+d}}, \quad (11)$$

for some universal  $C > 0$  and with probability at least  $1 - \frac{1}{n}$ . The above bound is in fact minimax optimal.

### 3.2 The $\delta$ -Separation and $\delta$ -Consistency Criteria

We begin by formulating a novel criterion of cluster separation that is naturally suited to smooth densities and is in fact equivalent to cluster separation in the merge distance of Eldridge et al. [2015]. See Section 5.2 below for details, as well as Kim et al. [2016].

**Definition 5.** *Two connected subsets  $A$  and  $A'$  of the support of the density  $p$  are  $\delta$ -separated when they belong to distinct connected components of the level set  $\{p > \lambda - \delta\}$ , where  $\lambda := \inf_{x \in A \cup A'} p(x)$ .*

The intuition behind the notion of  $\delta$ -separation is simple: the smoothness properties of the density limit the minimal degree of “vertical” and “horizontal” separation between clusters. This is illustrated in Figure 1 and best explained for the case of a density in  $\Sigma(\alpha, L)$  with  $\alpha \leq 1$ . If  $A$  and  $A'$  are  $\delta$ -separated, then their distance is at least  $(\frac{\delta}{L})^{1/\alpha}$ . And similarly, if two clusters  $A$  and  $A'$  are at a distance  $\sigma$  from each other (that is  $\sigma = \inf_{x \in A, y \in A'} \|x - y\|$ ), they are  $\delta$ -separated with  $\delta$  upper bounded by the same amount. As a result, separation between clusters of smooth densities can be defined using only one parameter, a feature that we will exploit to derive a new notion of consistency for clustering.

**Definition 6** ( $\delta$ -consistency). *Let  $\delta > 0$  and  $\gamma \in (0, 1)$ . A cluster tree estimator based on an i.i.d. sample  $\{X_i\}_{i=1}^n$  is  $(\delta, \gamma)$ -accurate if, with probability no smaller than  $1 - \gamma$ , for any pair of connected subsets  $A$  and  $A'$  of the support of  $p$  that are  $\delta$ -separated, exactly one of the following conditions holds:*

1. *at least one of  $A \cap \{X_i\}_{i=1}^n$  and  $A' \cap \{X_i\}_{i=1}^n$  is empty;*
2. *the smallest clusters in the cluster tree estimator containing  $A \cap \{X_i\}_{i=1}^n$  and  $A' \cap \{X_i\}_{i=1}^n$  are disjoint.*

*Let  $\{\delta_n\}$  be a vanishing sequence of positive numbers and a  $\{\gamma_n\}$  a vanishing sequence in  $(0, 1)$ . The sequence of cluster tree estimators  $\{\mathcal{T}_n\}_n$ , where  $\mathcal{T}_n$  is based on an i.i.d. sample  $\{X_i\}_{i=1}^n$  from  $p$ , is  $\delta$ -consistent with rate  $(\delta_n, \gamma_n)$  if, for all  $n$ ,  $\mathcal{T}_n$  is  $(\delta_n, \gamma_n)$ -accurate.*

The first condition in definition 6 is to rule out the trivial cases.

It is important to realize that the notion of  $\delta$ -consistency is a *uniform* notion of consistency: it is required to hold simultaneously over all possible pairs of  $\delta_n$ -separated connected subsets of the support, for an appropriate sequence  $\{\delta_n\}$ .

**Remark 2.** *For simplicity below we will take  $\gamma_n = \frac{1}{n}$ . It is of course possible to take  $\gamma = n^{-c}$  for any  $c > 0$ ; this will affect  $\delta_n$  only in the constants.*

### 3.3 The Split Levels

One of the most important features of the cluster tree of a density is the collections of levels  $\lambda$  at which the clusters split into two or more disjoint sub-clusters, which we refer to as *split levels*. Such levels correspond to critical changes in the topology of the upper level sets of  $p$ , of which clustering is a manifestation. One would hope that the split levels of a consistent cluster tree estimator should closely match the split levels of the cluster tree  $T_p$ . In what follows, we give a rigorous definition of the split levels and relate it to the criterion of  $\delta$ -separation of clusters. The notion of split levels will be important below in Section 3.4.2 in formalizing sufficient conditions

under which computationally efficient and statistically optimal cluster tree estimation is feasible for Hölder densities with smoothness degree  $\alpha$  greater than 1. It will also be used to demonstrate that our algorithms for cluster tree estimations are not only  $\delta$ -consistent, but will also not produce spurious clusters, with high probability (see Section 3.7).

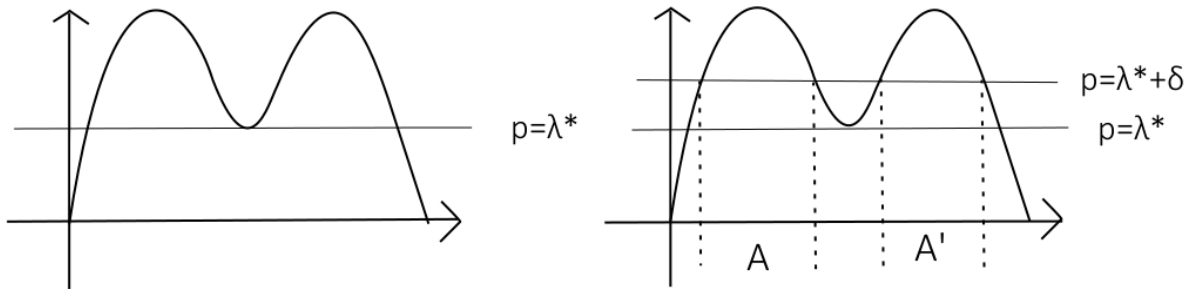


Figure 1: The left figure depicts a split level  $\lambda^*$  (see Definition 7) of the density  $p$ . The right figure depicts two  $\delta$ -separated sets  $A$  and  $A'$  with respect to  $\lambda^*$ .

**Definition 7.** Let  $p : \mathbb{R}^d \rightarrow \mathbb{R}$  is a continuous density. For fixed  $\lambda^* > 0$ , let  $\{\mathcal{C}_k\}_{k=1}^K$  be the collection of connected components of  $\{p \geq \lambda^*\}$ . The value  $\lambda^*$  is said to be a split level of  $p$  if there exists a  $\mathcal{C}_k$  such that  $\mathcal{C}_k \cap \{p > \lambda^*\}$  has two or more connected components.

The following, simple result illustrates the main topological properties of split levels.

**Proposition 2.** Suppose  $p : \mathbb{R}^d \rightarrow \mathbb{R}$  is compactly supported and that  $A$  and  $A'$  are subsets of two distinct connected components of  $\{p \geq \lambda_1\}$ . If  $A$  and  $A'$  belongs to the same connected components of  $\{p \geq \lambda_2\}$ , where  $\lambda_2 < \lambda_1$ , then there is a unique split level  $\lambda^* \in [\lambda_2, \lambda_1)$  such that  $A$  and  $A'$  belong to one connected component of  $\{p \geq \lambda^*\}$  and to distinct connected components of  $\{p > \lambda^*\}$ .

Proposition 2 suggests that if two connected components merge into one as the density level  $\lambda$  decreases, then there exists one and only one split level at which the corresponding merge takes place. Therefore, it is natural to make the following definition, which characterizes the corresponding split level of any two distinct clusters in a cluster tree.

**Definition 8.** Suppose  $A$  and  $A'$  are two open sets. Then  $A$  and  $A'$  are said to split at level  $\lambda^*$  if  $A$  and  $A'$  belong to one connected component of  $\{p \geq \lambda^*\}$  and to two distinct connected components of  $\{p > \lambda^*\}$ .

Furthermore, there is a direct link between the criterion of  $\delta$ -separated sets and the notion of split levels, as illustrated in the next result. This fact we will exploited later on in Section 3.7 to prove that the cluster tree estimators considered here also automatically yield accurate estimates of the splits levels and, therefore, do not lead to spurious clusters.

**Corollary 3.** *Let  $A$  and  $A'$  be  $\delta$ -separated. Then there exists a split level  $\lambda^*$  of the density, with*

$$\lambda^* \leq \inf_{x \in A \cup A'} f(x) - \delta,$$

*such that  $A$  and  $A'$  belong to one connected component of  $\{p \geq \lambda^*\}$  and to two connected components of  $\{p > \lambda^*\}$ .*

### 3.4 Rate of Consistency for the DBSCAN Algorithm

In this section, we will present the main results of the paper, and derive rates of consistency for DBSCAN-based cluster tree estimators of Hölder smooth densities in  $\Sigma(L, \alpha)$  with respect to the notion of  $\delta$ -separation. Specifically we show that these estimators are  $\delta$ -consistent with rate (see Definition 6 above)

$$\delta_n \geq Cn^{-\frac{\alpha}{2\alpha+d}}, \quad (12)$$

for an appropriate positive constant  $C$  that depend on  $\|p\|_\infty$ ,  $L$  and, possibly,  $K$  and  $\alpha$ . The above rates depend on the smoothness of the underlying density, with smoother densities leading to faster rates, and, as we prove later on in Section 3.6, are in fact minimax optimal. This is one of the main findings of the article and provides a sharpening over the consistency results of Chaudhuri et al. [2014] for cluster tree estimation, which are independent of the smoothness of  $p$ . As remarked above, (12) matches the optimal rate for density estimation given in (11).

We will carry out separate analyses for the case of  $\alpha \leq 1$  and the more subtle case in which the density has a higher degree of smoothness, i.e.  $\alpha > 1$ .

#### 3.4.1 Consistency for $\alpha \leq 1$

For Hölder densities with smoothness parameter  $\alpha \leq 1$ , the DBSCAN estimator given in Algorithm 1, which relies on the spherical kernel, is optimal. This result should not be particularly surprising, as Sriperumbudur and Steinwart [2012] have already demonstrated, using settings different from ours, that DBSCAN can be used optimally for density-based clustering. For completeness, we provide the proof of the consistency results.

In order to demonstrate that DBSCAN is  $\delta$ -consistent, it will be sufficient to show that the procedure provides an adequate approximation to the upper level sets of  $\hat{p}_h$ .

**Lemma 4.** *Assume that  $p \in \Sigma(\alpha, L)$ , where  $\alpha \in (0, 1]$ , and let  $K$  be the spherical kernel. Then, setting  $h = C_1 n^{-\frac{1}{2\alpha+d}}$ , for any  $C_1 > 0$ , there exist a constant  $C_2 > 0$ , depending on  $C_1$ ,  $\|p\|_\infty$ ,  $L$  and  $d$  such that, uniformly over all  $\lambda > 0$ , with probability at least  $1 - \frac{1}{n}$ ,*

$$\left\{ p \geq \lambda + C_2 \frac{\log(n)}{n^{\alpha/(2\alpha+d)}} + Lh^\alpha \right\} \subset \bigcup_{X_j \in \hat{D}(\lambda)} B(X_j, h) \subset \left\{ p \geq \lambda - C_2 \frac{\log(n)}{n^{\alpha/(2\alpha+d)}} - Lh^\alpha \right\}. \quad (13)$$

As a direct corollary of Lemma 4, we see that setting  $h$  to be of order  $n^{-\frac{1}{2\alpha+d}}$ , then the DBSCAN algorithm will output a  $\delta$ -consistent cluster tree at rates that are adaptive to  $\alpha$ .

**Corollary 5.** *Under the assumptions of Lemma 4, the cluster tree returned by the DBSCAN Algorithm 1 is  $\delta$ -consistent with rate  $\delta_n \geq C \frac{\log n}{n^{\alpha/(2\alpha+d)}}$ , where  $C = C(\|p\|_\infty, L, d)$  is a constant independent of  $\delta$ .*

### 3.4.2 Consistency for $\alpha > 1$

When  $\alpha > 1$ , the vanilla Algorithm 1 no longer delivers the optimal rate (12), for statistical and computational reasons. The statistical reasons are clear: when  $\alpha > 1$  it become necessary to rely on smoother kernels, namely  $\lceil \alpha \rceil$ -valid kernels as indicated Section 3.1. This will lead to a bias  $\|p - p_h\|_\infty$  of the correct order  $O(h^\alpha)$  and, therefore, by choosing the bandwidth as in (10), to the optimal balance in the bias/variance trade-off as in (11). The computational reasons are more subtle: in order to determine cluster connectedness DBSCAN employs single-linkage type rules, which will force a sub-optimal choice for the bandwidth even if the kernel  $K$  is chosen to be  $\lceil \alpha \rceil$ -valid. To exemplify, suppose we would like to cluster the sample points  $X_{i_1}, \dots, X_{i_k}$  belonging the upper level set  $\{x: \hat{p}_h(x) \geq \lambda\}$ , for some  $\lambda > 0$ . The computationally efficient, single-linkage rule implemented by DBSCAN for a choice of the input  $h$  is to clusters the points based on the connected components of the union-of-balls around them, i.e based on the connected components of

$$\hat{L}(\lambda) = \bigcup_{j=1}^k B(X_{i_j}, h).$$

Assume now that the gradient of  $p$  has norm uniformly bounded by a constant  $D$  for all  $x \in L(\lambda)$ . Then,

$$\max_{j=1, \dots, k} \sup_{x \in B(X_{i_j}, h)} |p(x) - p(X_{i_j})| \leq Dh, \quad (14)$$

and, as a result,

$$\left\{ p \geq \lambda + C \left( \frac{\log(n)}{\sqrt{nh^d}} + h^\alpha \right) + Dh \right\} \subset \hat{L}(\lambda) \subset \left\{ p \geq \lambda - C \left( \frac{\log(n)}{\sqrt{nh^d}} - h^\alpha \right) - Dh \right\}, \quad (15)$$

where the terms  $C \left( \frac{\log(n)}{\sqrt{nh^d}} - h^\alpha \right)$  come from the  $L_\infty$  error bound of  $\lceil \alpha \rceil$ -valid kernel as in (9) and  $Dh$  is due to of (14). As  $h \rightarrow 0$ , the term  $Dh$  dominates the bias term, of order  $O(h^\alpha)$ , so that the optimal choice of  $h$  is of the order  $h \asymp \frac{\log n}{2+d}$ , which in turn yields a worse rate than (12). What is more, on the event that

$$\min_{j=1, \dots, k} \|\nabla p(X_{i_j})\| \geq D',$$

for some  $D' > 0$ , we have that, as  $h \rightarrow 0$ ,  $\sup_{x \in B(X_{i_j}, h)} |p(x) - p(X_{i_j})| = \Theta(h)$  for each  $j$ . Then, on that event, the inclusions in (15) are tight, showing that the sub-optimal choice of  $h$  cannot be ruled out.

We believe that these considerations, which reveal as an interesting trade-off between computationally efficiency and statistical optimality, apply to not just DBSCAN, but to single-linkage type of algorithms.

The issue outline above can be handled in more than one way. A neraly trivial but impractical solution, studied in detal in Section 3.5below, is to deploy a computationally inefficient algorithm that assumes the ability to evaluate the connected components of the upper level set of  $\hat{p}_h$  exactly: see Algorithm 3 below. It is very easy to see that this will result in optimal  $\delta$ -consistency (see Corollary 8 below). Unfortunately, this procedure will require evaluating  $\hat{p}_h$  on a very fine grid, a task that is computationally unfeasible even in small dimensions. The second, more interesting and novel approach, which we describe next, is to further assume that  $p$  satisfies mild additional regularity conditions around the split levels that are reminiscent of low-noise type assumptions in

classification. Under those assumptions, the modified DBSCAN Algorithm 2, given below, will achieve the optimal rate (12) while remaining computationally feasible since it only operates on the sample points.

---

**Algorithm 2** The modified DBSCAN

---

**INPUT:** i.i.d sample  $\{X_i\}_{i=1}^n$ , a  $\lceil\alpha\rceil$ -valid kernel  $K$  and  $h > 0$

1. Compute  $\{\hat{p}_h(X_i), i = 1, \dots, n\}$ .

For each  $\lambda \geq 0$ ,

2. construct a graph  $\mathbb{G}_{h,\lambda}$  with node set

$$\hat{D}(\lambda) = \{X_i : \hat{p}_h(X_i) \geq \lambda\}$$

and edge set  $\{(X_i, X_j) : X_i, X_j \in \hat{D}(\lambda) \text{ and } \|X_i - X_j\| \leq 2h\}$ .

3. Compute  $\mathbb{C}(h, \lambda)$ , the maximal connected components of  $\mathbb{G}_{h,\lambda}$ .

**OUTPUT:**  $\hat{T}_n = \{\mathbb{C}(h, \lambda), \lambda \geq 0\}$

---

**Remark 3.** *Despite its seemingly different form, Algorithm 2 is nearly identical to Algorithm 1. The only difference is in the use of a  $\lceil\alpha\rceil$ -valid kernel  $K$  instead of a spherical kernel. Furthermore, the procedures only requires evaluating at most  $n + 1$  different graphs:*

$$\mathbb{G}_{h,0}, \mathbb{G}_{h,\hat{p}_h(X_{\sigma_1})}, \dots, \mathbb{G}_{h,\hat{p}_h(X_{\sigma_n})},$$

where  $(\sigma_1, \dots, \sigma_n)$  is a permutation of  $(1, \dots, n)$  such that

$$\hat{p}_h(X_{\sigma_1}) \leq \hat{p}_h(X_{\sigma_2}) \leq \dots \leq \hat{p}_h(X_{\sigma_n})$$

And, again just like with Algorithm 1, the connected components of each  $\mathbb{C}(h, \lambda)$  can be easily evaluated by maintaining a union-find structure.

We will now describe the extra regularity conditions we will impose on the geometry of the density  $p \in \Sigma(\alpha, L)$  around the split levels that guarantees optimality of the clustering Algorithm 2. We begin by formulating two widely used technical conditions on a generic set  $\Omega \subset \mathbb{R}^d$ .

**C1.** (The Standard Assumption) There exist constants  $r_I, c_I > 0$  such that, for any  $0 \leq r \leq r_I$  and  $x \in \Omega$ ,

$$\mathcal{L}(B(x, r) \cap \Omega) \geq c_I V_d r^d,$$

where we recall that  $\mathcal{L}$  here denote the Lebesgue measure of  $\mathbb{R}^d$ .

**C2.** (The Covering Condition) There exists a constant  $C_I$  such that, for any  $0 < r$ , there exists a collection of points  $\mathcal{N}_r \subset \Omega$  such that  $\text{card}(\mathcal{N}_r) \leq C_I r^{-d}$  and

$$\bigcup_{y \in \mathcal{N}_r} B(y, r) \supset \Omega.$$

Conditions **C1**, **C2** hold for many sets  $\Omega$ . In particular, they hold for compact manifolds with piecewise Lipschitz boundary (see, e.g, Do Carmo [1992]). Since  $p \in \Sigma(L, \alpha > 1)$ , any upper-level set  $\{p \geq \lambda\}$  is a union of connected  $d$  dimensional manifolds with  $C^1$  boundary and therefore

meet both **C1** and **C2**. Condition **C1** is known as the inner cone condition **C1** in [Korostelev and Tsybakov \[1993\]](#); the term standard condition is due to [Cuevas \[2009\]](#).

We will require both **C1** and **C2** to hold simultaneously for all the upper level-sets of  $p$  right above the split levels. Specifically, we will assume that

**C.** There exists a  $\delta_0 > 0$  such that, for any split level  $\lambda^*$  of  $p$  and any  $0 < \delta \leq \delta_0$ , the set  $\{x: p(x) \geq \lambda^* + \delta\}$  satisfies conditions **C1** and **C2** with universal constant  $r_{I,C_I}$  and  $C_I$  only depending on  $p$ .

We also need the connected components of the upper level sets right above split levels to be sufficiently well separated in the . Below we introduce another condition that essentially characterizes the separation of the distinct connected components right above the split levels.

**S( $\alpha$ ).** There exist positive constants  $\delta_S$  and  $c_S$  such that, for each split level  $\lambda^*$  of  $p$ , the following holds: let  $\{\mathcal{C}_k\}_{k=1}^K$  be the connected components of  $\{x: p(x) > \lambda^*\}$ . Then,

$$\min_{k \neq k'} d(\mathcal{C}_k \cap \{p \geq \lambda^* + \delta\}, \mathcal{C}_{k'} \cap \{p \geq \lambda^* + \delta\}) \geq c_S \delta^{1/\alpha}, \quad \forall \delta \in (0, \delta_S]. \quad (16)$$

Our following result shows that [Corollary 5](#) still holds for  $\alpha > 1$  .

**Theorem 6.** *Let  $p \in \Sigma(\alpha > 1, L)$  be any density function with compact connected support and finitely many split levels bounded from below by  $\lambda_0 > 0$ . Assume also that conditions **C** and **S( $\alpha$ )** hold for  $p$ . Then, the modified DBSCAN Algorithm [2](#) is  $(\delta_n, \gamma_n)$  consistent, where  $\gamma_n = \frac{1}{n} + O(h^{-d} \exp(-c\lambda_0 n^{\alpha/(2\alpha+d)}))$ ,  $c$  is a positive constant only depending on  $p$  and*

$$\delta = 2a_n + (4h/c_S)^\alpha$$

with  $a_n$  defined in [\(9\)](#) and  $c_S$  the constant in **C**. Thus if  $h \asymp n^{-1/(2\alpha+d)}$ , the cluster tree returned by the modified DBSCAN algorithm is consistent with  $\delta = \Omega(n^{-\alpha/(2\alpha+d)})$  with high probability.

The explicit expression of  $c$  in the proposition can be found in [\(35\)](#). The proof of the theorem heavily relies on both condition conditions **C** and **S( $\alpha$ )**, which cannot be dispensed of. Luckily, both conditions hold for a variety of densities, as we show in the next two examples.

**Example 1: Morse densities.** We recall that a function  $p$  on  $\mathbb{R}^d$  is Morse if all the critical points of  $p$  have a non-degenerate Hessian. See, e.g., [Matsumoto \[2002\]](#) for an treatment of Morse theory. An equivalent and more intuitive condition is that around the critical points,  $p$  behaves like a quadratic function. The assumption that a density is Morse is routinely used in the literature on density-based clustering and mode estimation: see, e.g., [Chacón et al. \[2015\]](#) and [Arias-Castro et al. \[2016\]](#) and references therein. In the [Proposition 19](#) of the appendix, we show that any Morse function on  $\mathbb{R}^d$  satisfies **C** and **S(2)**.

**Example 2: Natural Splines.** Let  $\alpha \geq 2$  be any integer. Let  $f_1 : [1, 2] \rightarrow \mathbb{R}$  be such that  $f_1(x) = (x - 2)^\alpha$ . It is easy to find a polynomial  $f_2$  of degree  $\alpha$  on  $[0, 1]$  such that the real valued function  $f$  on  $\mathbb{R}_+$  given by

$$f(x) = \begin{cases} f_1(x), & x \in [1, 2] \\ f_2(x), & x \in [0, 1] \\ 0, & \text{otherwise,} \end{cases}$$

has continuous derivatives up to order  $\alpha - 1$  and is such that  $f(0) = f'(0) = \dots = f^{(\alpha-1)}(0) = 0$ . When  $\alpha = 3$ ,  $f$  is called a natural spline. For any dimension  $d$ , let  $F : \mathbb{R}^d \rightarrow \mathbb{R}$  be such that  $F(x) = f(\|x\|)$ . Then,  $F \in \Sigma(\alpha, L)$ . Denote  $x_0 = (2, 0, \dots, 0)$ . Let  $G$  be the function  $x \in \mathbb{R}^d \mapsto F(x - x_0) + F(x + x_0)$ . It is easy to see that, for any  $0 < \delta \leq 1$ ,

$$\{G(x) \geq \delta\} = B(x_0, 2 - \delta^{1/\alpha}) \cup B(-x_0, 2 - \delta^{1/\alpha}).$$

As a result conditions **C** and **S**( $\alpha$ ) are trivially satisfied in this simple case. This example also implies that if the density locally behaves like a spline function of order  $\alpha$ , then conditions **C** and **S**( $\alpha$ ) are satisfied.

### 3.5 Consistency of the Density Cluster Tree of $\hat{p}_h$

As a side result, we show below that the cluster tree of the KDE  $\hat{p}_h$  is, for an appropriate choice of the bandwidth  $h$ , a minimax optimal estimator of the cluster tree of  $p$  with respect to the  $\delta$ -separation criterion. As remarked above, such an estimator, given below in Algorithm 3, is not computable even in small dimensions, as it requires evaluating the connected components of all the upper level sets of  $\hat{p}_h$ .

---

#### Algorithm 3 Clustering based on connected components

---

**INPUT:** i.i.d sample  $\{X_i\}_{i=1}^n$ , the kernel  $K : \mathbb{R}^d \rightarrow \mathbb{R}$ , the level  $\lambda$  and  $h > 0$

1. Compute  $\hat{L}(\lambda) = \{x : \hat{p}_h(x) \geq \lambda\}$ .
2. Construct a graph  $\mathbb{G}_{h,k}$  with nodes

$$\hat{D}(\lambda) = \{X_i\}_{i=1}^n \cap \hat{L}(\lambda)$$

and edges  $(X_i, X_j)$  if  $X_i$  and  $X_j$  belong to the same connected component of  $\hat{L}(\lambda)$ .

3. Compute  $\mathbb{C}(h, \lambda)$ , the graphical connected components of  $\mathbb{G}_{h,\lambda}$ .

**OUTPUT:**  $\mathbb{C}(h, \lambda)$

---

We show that for generic  $\alpha > 0$ , if  $p \in \Sigma(L, \alpha)$ , level sets of KDE estimator are good approximations of the corresponding population quantities.

**Lemma 7.** *Assume that  $p \in \Sigma(L, \alpha)$ , where  $\alpha > 0$ , and let  $K$  be a  $[\alpha]$ -valid kernel. Suppose  $h = h_n = C_1 \left( \frac{1}{n^{1/(2\alpha+d)}} \right)$  for some absolute constant  $C_1$ . Then there exists  $C_2 > 0$ , depending on  $C_1, \|p\|_\infty, K, L$  and  $d$  such that, when  $nh^d \geq 1$ , with probability  $1 - 1/n$ , uniformly over all  $\lambda > 0$ ,*

$$\left\{ x : p(x) \geq \lambda + C_2 \frac{\log(n)}{n^{\alpha/(2\alpha+d)}} \right\} \subset \{x : \hat{p}_h(x) \geq \lambda\} \subset \left\{ x : p(x) \geq \lambda - C_2 \frac{\log(n)}{n^{\alpha/(2\alpha+d)}} \right\}, \quad (17)$$

As a direct corollary of Lemma 7 we can establish the following consistency rate for the algorithm 3.

**Corollary 8.** *Let  $h = h_n = C_1 \left( \frac{1}{n^{1/(2\alpha+d)}} \right)$ . Under the assumptions of Lemma 7, the cluster tree returned by Algorithm 3 is  $\delta$ -consistent with probability at least  $1 - \gamma$ , where*

$$\delta \geq C \frac{\log n}{n^{\alpha/(2\alpha+d)}}, \quad (18)$$

with  $C = C(\|p\|_\infty, K, L, \gamma, d)$  a constant independent of  $n$  and  $\delta$ .



We remark that while Algorithm 3 is not feasible even in moderate dimension, its optimality holds without any additional regularity conditions such as  $\mathbf{S}(\alpha)$  and  $\mathbf{C}$ .

### 3.6 Lower bounds

Below we show that the rates of the DBSCAN algorithm derived in the previous sections are minimax optimal. We remark that in our analysis we cannot use the lower bound construction of Chaudhuri et al. [2014], as we consider the sub-class of Hölder smoother densities. Another minor difference is that our result applies to all dimensions  $d$ , while the arguments in Chaudhuri et al. [2014] requires  $d \geq 2$ . We recall that the notion of  $\delta$ -accuracy is given in Definition 6.

**Lemma 9.** *Suppose for fixed  $d \geq 1$  and  $\alpha > 0$ . There exists a finite family  $\mathcal{F}$  of  $d$ -dimensional probability density functions belonging to the Hölder class  $\Sigma(L, \alpha)$  satisfying  $\mathbf{C}$  and  $\mathbf{S}(\alpha)$  and uniformly bounded from above by  $C_0$ , and a constant  $\mathcal{K}$ , depending on  $L$  and  $\alpha$ , such that the following holds when*

$$n \geq \frac{4^d 8 \log(32)}{V_d} \quad \text{and} \quad \delta \leq \min \left\{ \left( \frac{\mathcal{K}}{16^\alpha (7C_0)^{\alpha/d}} \right), \|p\|_\infty / (2^{d/2+1}) \right\},$$

where  $V_d$  denote the volume of a  $d$  dimensional ball. If a cluster tree estimator of  $p$  is  $(\delta, 1/4)$ -accurate when presented with an i.i.d. sample  $\{X_i\}_{i=1}^n$  from a density in  $\mathcal{F}$ , then it must be the case that

$$n \geq \frac{\|p\|_\infty \mathcal{K}^{d/\alpha}}{C \delta^{2+d/\alpha}}, \quad (19)$$

for some constant  $C$  only dependent on  $d$ .

It is important to remark that the class of functions  $\mathcal{F} \subset \Sigma(\alpha, L)$  used in the proof of Lemma 9 also satisfies conditions  $\mathbf{C}$  and  $\mathbf{S}(\alpha)$ . See Lemma 20 in B shows for more details. Since the lower bound in Lemma 9 is of the same order as the upper bound from Theorem 6 the consistency rate established in that result is minimax optimal.

**Remark 4.** *Assuming  $\|p\|_\infty$  and  $d$  fixed and  $\gamma$  logarithmic in  $n$ , the cluster consistency guarantees of the DBSCAN-based algorithms derived in Corollary 5 and Theorem 6 and the lower bound in (19) differ only in the constants and by a term of order  $\log n$ . Thus, altogether these results show that, up to a log factor, the optimal clustering rate for  $\delta$ -consistency for density functions in  $\Sigma(L, \alpha)$  is  $\Theta\left(\frac{\log(n)}{n^{\alpha/(2\alpha+d)}}\right)$ , achieved by Algorithm 1 when  $\alpha \leq 1$  and Algorithm 2 when  $\alpha > 1$ . Interestingly, up to logarithmic terms, the rate matches the minimax rate for estimating  $\alpha$ -smooth densities in the  $L_\infty$  norm: see Korostelev and Nussbaum [1999]. In hindsight, this finding is not very surprising. Indeed, as noted in the discussion,  $\delta$ -separation of clusters is equivalent to separation in the merge distance of Eldridge et al. [2015] (see Definition 12 below), which, in turn, as shown in Kim et al. [2016], can be linked to the supreme norm of the difference between  $p$  and  $\hat{p}$ , if  $p$  is continuous. Thus, for continuous densities, the performance of a cluster tree estimator based on a density estimator  $\hat{p}$  ought to be tied to  $\|p - \hat{p}\|_\infty$ . Though a similar fact was also noted in Chaudhuri and Dasgupta [2010], this connection has not been previously established in the literature in a rigorous manner.*

### 3.7 Consistent estimate of the Split levels

We now discuss a simple pruning procedure for the cluster tree estimators considered here that allows to estimate consistently the split levels of the underlying density and, as a result, is guaranteed to prevent the occurrence of spurious estimated clusters. Pruning and consistent estimation of split levels have been previously analyzed in [Sriperumbudur and Steinwart \[2012\]](#) when  $\alpha \leq 1$  and in [Chaudhuri et al. \[2014\]](#) for general densities. However, none of the existing algorithms give error bounds being adaptive to  $\alpha$  for density  $p \in \Sigma(\alpha, L)$  with  $\alpha > 1$ .

The following definition provides a way to identify significant split levels in the cluster tree estimator returned by Algorithm 2.

**Definition 9.** *Given  $\Delta > 0$ , the value  $\widehat{\lambda}^* \in (0, \infty)$  is said to be a  $\Delta$ -significant split level of the cluster tree estimator if there exist two data points  $X_i, X_j \in D(\widehat{\lambda}^* + \Delta)$  satisfying*

$$\widehat{\lambda}^* = \sup\{\lambda > 0 : X_i \text{ and } X_j \text{ are in the same connected component of } \mathbb{C}(h, \lambda).\} \quad (20)$$

Thus a  $\Delta$ -significant split level of the cluster tree estimator is a split level such that the clusters “born” at that level persists also at higher levels.

The intuition of the  $\Delta$ -significant split level is that in theory, the accuracy of modified DBSCAN estimator is limited with finitely many data points. By looking at split levels corresponding to large clusters, we rule out the insignificant split levels and only keep the  $\Delta$ -significant ones. Therefore finding  $\Delta$ -significant levels can be thought of as a process of pruning the cluster tree estimators.

Below, we show that there is a one to one correspondence between  $\Delta$ -significant split level of the modified DBSCAN cluster tree estimator and the split level of the population density under a slightly stronger covering condition than **C**:

**C’.** There exists  $\delta_0 > 0$  such that for any split level  $\lambda^*$  of  $p$  and any  $|\delta| \leq \delta_0$ ,  $\{p \geq \lambda^* + \delta\}$  satisfies conditions **C1** and **C2** with universal constant  $r_{I, C_I}$  and  $C_I$  only depending on  $p$ .

The only difference between **C** and **C’** is that while condition **C** assumes the regularity above split levels, **C’** ensures the regularity around split levels.

**Proposition 10.** *Suppose condition **C’** and **S** hold. Let  $\Delta = 2a_n + (4h/c_S)^\alpha$  where  $a_n = \frac{C(\gamma + \log(1/h))}{\sqrt{nh^d}} + C(L, K, \alpha)h^\alpha$  is defined in (9) and  $h = C_1 n^{-1/(2\alpha+d)}$ . Suppose  $p$  has finitely many split levels. Then, with probability at least  $1 - 1/n - O(h^{-d} \exp(-cn^{\alpha/(2\alpha+d)}))$ , the following additional results hold:*

**1.** *Let  $\lambda^*$  be a split level of the density  $p$ . Suppose  $\mathcal{C}$  and  $\mathcal{C}'$  are two open sets splitting at  $\lambda^*$  (see Definition 8) and that*

$$\min\{P(\mathcal{C} \cap \{p \geq \lambda^* + 2\Delta\}), P(\mathcal{C}' \cap \{p \geq \lambda^* + 2\Delta\})\} > 0 \quad (21)$$

*Then there exists a split level  $\widehat{\lambda}^*$  of the cluster tree estimator (constructed by the modified DBSCAN) being  $\Delta$ -significant such that*

$$|\lambda^* - \widehat{\lambda}^*| \leq \Delta \quad (22)$$

**2.** *Conversely suppose  $\widehat{\lambda}^*$  is a  $\Delta$ -significant split level of the cluster tree estimator. Then there exists a split level  $\lambda^*$  of  $p$  such that*

$$|\lambda^* - \widehat{\lambda}^*| \leq \Delta. \quad (23)$$

Proposition 10 says that with high probability, every  $\Delta$ -split level corresponds to a density split level and that conversely, any split level of the density  $p$  can be found if we have enough data. To prune the cluster tree returned by the modified DBSCAN algorithm, it suffices to remove all the split levels that are not  $\Delta$ -significant.

## 4 Densities with Gaps

In this final section we investigate the properties of DBSCAN-based cluster tree estimators when the underlying density  $p$  is no longer continuous but exhibit instead a jump discontinuity, so that, for all values of  $\lambda$  in a given interval of length  $\epsilon$ , the upper level sets  $\{p \geq \lambda\}$  do not change. The value of  $\epsilon$  is referred to as the gap size. See Figure 2 for examples.

We formally define the notion of density gap below.

**Definition 10.** A Lebesgue density  $p$  in  $\mathbb{R}^d$  is said to have a gap of size  $\epsilon > 0$  at level  $\lambda_* > 0$  when

$$\inf_{x \in S} p(x) \geq \lambda^* \quad \text{and} \quad \sup_{x \in S^c} p(x) \leq \lambda_*,$$

where  $\lambda^* = \lambda_* + \epsilon$  and

$$S = \{p(x) \geq \lambda^*\}. \tag{24}$$

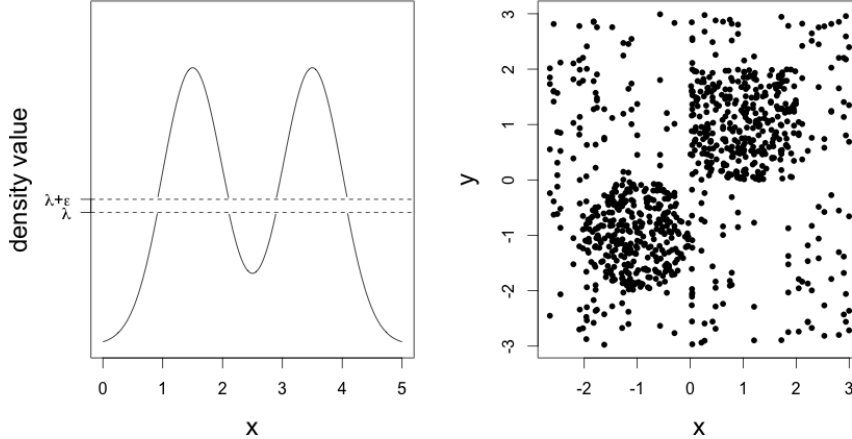


Figure 2: The left plot depicts a one dimensional density with gap of size  $\epsilon$  at level  $\lambda$ . It is clear that  $\{\lambda < p < \lambda + \epsilon\}$  is an empty set. The right plot depicts 500 i.i.d sampling from a two dimensional density with a gap. It is clear that the density is low in the background and high on the disk centered at  $(-1,-1)$  and on the square centered at  $(1,1)$ . Finding the samples points with low density values can be thought of as outliers detection in this case.

It follows from the above definition that, if  $p$  has a gap of size  $\epsilon$  at  $\lambda_*$ , then

$$L(\lambda) = \bigcup_{i=1}^I \mathcal{C}_i \equiv S, \quad \forall \lambda \in (\lambda_*, \lambda_*^*],$$

where  $(\mathcal{C}_1, \dots, \mathcal{C}_I)$  are disjoint, connected sets in  $\mathbb{R}^d$ . To avoid trivialities, we will assume throughout that  $I \geq 1$ . Though fairly restrictive, this scenario is already interesting for the purpose of both clustering and level set estimation. Indeed, this situation encompasses the ideal clustering scenario, depicted as examples in Figures 1 and 5 in the original DBSCAN paper Ester et al. [1996], of a piecewise constant density that is low everywhere on its support with the exception of a few connected, full-dimensional regions, or clusters, where it is higher by a certain amount (in our case the gap  $\epsilon$ ). The size of the gap parameter  $\epsilon$  and the minimal distance among clusters both affect the difficulty of the clustering task, which becomes harder as both parameters get smaller. In our analysis, we keep track of such dependence, thus producing consistency rates depending on the sample size, the gap size and the minimal distance. In particular, we will demonstrate that DBSCAN algorithm achieves the optimal minimax scaling in both parameters.

#### 4.1 Connection to the Devroye-Wise Estimator and Minimax Optimal Estimation of Level Sets

In the simplified setting considered here, it turns out that the DBSCAN algorithm is equivalent to the renown Devroye-Wise estimator of the level set  $\{x: p(x) \geq \lambda^*\}$ ; see Devroye and Wise [1980]. Below we formalize this observation and show that this estimator is minimax optimal for estimating the level set  $S$  when the gap size  $\epsilon$  is vanishing in the sample size  $n$ , in the sense that it allows the fastest possible decay of  $\epsilon$ . To the best of our knowledge, such scaling has not been previously established.

Recall that with inputs  $k$  and  $h$ , the DBSCAN algorithm outputs a set of nodes  $\mathbb{G}_{h,k}$ . One then may construct the random set

$$\widehat{S}_h = \bigcup_{X_j \in \mathbb{G}_{h,k}} B(X_j, h) \tag{25}$$

comprised of the union of balls of radius  $h$  around such points, and use it as an estimator for the high density region  $S = \cup_i \mathcal{C}_i$  consisting of all the clusters. The above estimator was originally proposed by Devroye and Wise [1980] to estimate the support of the underlying density.

We measure the performance of any estimator  $\widehat{S}$  with the Lebesgue measure of its symmetric difference with  $S$ :

$$\mathcal{L}(S \Delta \widehat{S}) = \mathcal{L}(S \cap \widehat{S}^c) + \mathcal{L}(S^c \cap \widehat{S}).$$

In order to determine the difficulty of this estimation problem we will need to use the gap size  $\epsilon$  as a parameter which may depend on  $n$ . We will in addition impose the following condition:

**(R).** (Level set regularity). There exists constants  $h_0 > 0$  and  $C_0 > 0$  such that, for all  $h \in (0, h_0)$ ,

$$\mathcal{L}(S_h \setminus S_{-h}) \leq C_0 h,$$

where  $S$  is as in (24) and  $S_h$  and  $S_{-h}$  are defined in (1).

The condition **R** is a very mild assumption. In particular, every compact domain in  $\mathbb{R}^d$  with  $C^2$  boundary satisfies condition **R**. In this case  $C_0 \leq V_{d-1}|\partial S|$ , where  $|\partial S|$  denotes the surface volume of  $S$ , and  $h_0$  is the size of the tubular neighborhood of  $\partial S$ .

**Proposition 11.** *Assume that the density  $p$  has gap of size  $\epsilon$  at level  $\lambda_*$  and that condition **(R)** holds. Let  $a_n = \frac{C(\gamma+\log(1/h))}{\sqrt{nh^d}}$  be defined as in (7). Suppose the input parameters  $(h, k)$  of the DBSCAN Algorithm 1 satisfy*

$$h_0 \geq h \geq C_1 \left( \frac{\log(n)}{n\epsilon^2} \right)^{1/d}, \quad (26)$$

where  $C_1 > 0$  depends on  $C$  and  $d$  and  $h_0$  is the constant appearing in assumption **R**, and

$$k = \lceil nh^d V_d \lambda \rceil,$$

where  $\lambda \in (\lambda_* + a_n, \lambda_* - a_n]$ . Then with probability at least  $1 - e^{-\gamma}$ ,

$$\mathcal{L}(S \Delta \widehat{S}_h) \leq 2C_0 h,$$

where  $C_0$  is the constant appearing in condition **(R)** and

$$\widehat{S}_h = \bigcup_{X_j \in \mathbb{G}_{h,k}} B(X_j, h). \quad (27)$$

If  $P(S) = 1$ , the level set estimator  $\widehat{S}_h$  is also a support estimator and  $\epsilon = \inf_{x \in S} p(x)$ . In this case, Proposition 11 says that if the lower bound of the density goes to 0 not faster than  $O(n^{-1/2})$ , then support estimation is still possible.

Below we show that the error bound given in Proposition 11 is minimax optimal up to log factors. Consider the collection  $\mathcal{P}^n(h_0, \epsilon)$  of probability distributions of  $n$  i.i.d. random vectors in  $\mathbb{R}^d$  from a distribution with bounded  $d$ -dimensional Lebesgue densities having a gap size  $\epsilon$  at some level and satisfying condition **(R)** with parameter  $h_0 > 0$ . Then Proposition 11 shows that

$$\sup_{P \in \mathcal{P}^n(h_0, \epsilon)} \mathbb{E}_P \left( S \Delta \widehat{S}_h \right) = O \left( \left( \frac{\log(n)}{n\epsilon^2} \right)^{1/d} \right),$$

provided that  $h$  is of the order  $\left( \frac{\log(n)}{n\epsilon^2} \right)^{1/d}$ .

In our next result we construct obtain a matching lower bound (up to a log factor).

**Proposition 12.** *There exists a  $h_0 > 0$  and constant  $c > 0$ , depending on  $d$  only such that for any  $\epsilon < 1/4$ ,*

$$\inf_{\widehat{S}} \sup_{P \in \mathcal{P}^n(h_0, \epsilon)} \mathbb{E}_P \left( S \Delta \widehat{S}_h \right) \geq c \min \left\{ \left( \frac{1}{n\epsilon^2} \right)^{1/d}, 1 \right\},$$

where the infimum is with respect to all estimators of  $S$ .

We remark that if  $n\epsilon^2 \rightarrow 0$  as  $n \rightarrow \infty$ , then Proposition 11 and Proposition 12 match up to a log factor of  $n$ . In other words, with suitable choice of input, DBSCAN can optimally estimate the level set  $S$  at the gap.

**Remark 5.** *Our results offer a sharpening of the rates of consistency of the Devroye-Wise estimator given in Cuevas and Rodríguez-Casal [2004] (see, e.g. Theorem 4 and 5 therein) where the size of the gap  $\epsilon$  is assumed fixed. Instead, we derive the optimal minimax scaling  $\epsilon$ ,  $w$  and show that the Devroye-Wise estimator, with an appropriate choice of  $h$  can achieve it. To the best of our knowledge, this result is new.*

## 4.2 Clustering at the Gap Level

We conclude this section by showing that DBSCAN estimates the clusters in  $S$  optimally.

In addition to assuming the existence of a gap of size  $\epsilon$ , we also need to measure the degree of separation among clusters in  $S$ .

**S**( $\sigma$ ). (Separation condition). There exists a constant  $\sigma > 0$  such that

$$\min_{i \neq j} \text{dist}(\mathcal{C}_i, \mathcal{C}_j) = \sigma.$$

The above condition is hardly an assumption: if the distance between two distinct cluster were 0 then the cluster may not be consistently estimated. We will let the parameters  $\sigma$  and  $\epsilon$  to vanish as  $n \rightarrow \infty$  and derive the minimax scaling for both of them. These minimax scaling describe the fastest rates of decay for the gap size and the separation parameter for which clustering is (barely) possible. In particular, we will demonstrate that the DBSCAN algorithm can optimally estimate the clusters at  $\lambda^*$  with suitable inputs under such scaling, implying that no other algorithm can do better than DBSCAN for such task.

**Proposition 13.** *Assume that  $p$  has a gap of size  $\epsilon > 0$  at level  $\lambda_* \geq 0$  and satisfy the separation condition **S**( $\sigma$ ). Suppose also that the input parameters  $(h, k)$  of the DBSCAN Algorithm 1 satisfy*

$$\sigma/4 \geq h \geq C_1 \left( \frac{\log(n)}{n\epsilon^2} \right)^{1/d} \quad \text{and} \quad k = \lceil nh^d V_d \lambda \rceil, \quad (28)$$

for some  $C_1$  and  $\lambda$  is any value in  $(\lambda_* + a_n, \lambda^* - a_n]$  where  $a_n = \frac{C(\gamma + \log(1/h))}{\sqrt{nh^d}}$  as in (7). Then the following results hold with probability at least  $1 - e^{-\gamma}$ :

- i. simultaneously over all connected sets  $A$  such that  $A_{2h} \subset \mathcal{C}_i$ , for some  $i$ , all the sample points in  $A$ , if any, belong to the same connected component of  $\mathbb{G}_{k,h}$ ;
- ii. simultaneously over all connected sets  $A$  and  $A'$  such that  $A_{2h} \subset \mathcal{C}_i$  and  $A'_{2h} \subset \mathcal{C}_j$ , for some  $i \neq j$ , the sample points in  $A$  and  $A'$ , if any, belong to distinct connected components of  $\mathbb{G}_{k,h}$ .

The definition of  $A_{2h}$  and  $A_{-2h}$  can be found in (1).

Proposition 13 implies the DBSCAN algorithm will yield clustering consistency provided (28) holds. This requires that the parameters  $\epsilon$  and  $\sigma$  are such that

$$n \geq C \frac{1}{\epsilon^2 \sigma^d},$$

for some constant  $C$ . The above inequality constrains the rate of decay of  $\epsilon$  and  $\sigma$  in terms of  $n$ . As it turns out, the resulting scaling is in fact minimax optimal, in the sense that any sequences of values for  $\epsilon$  and  $\sigma$  such that  $n = o\left(\frac{1}{\epsilon^2 \sigma^d}\right)$  will lead to inconsistent clustering, no matter the clustering algorithm used.

**Proposition 14.** *Consider a finite family of density functions  $F = \{f_j\}$ . Suppose all  $f_j \in F$  have gap of size  $\epsilon > 0$  at level  $\lambda_*$ . This means that for any  $j$ ,  $\{f_j \geq \lambda_* + \epsilon\} \cup \{f_j \leq \lambda_*\} = \mathbb{R}^d$ . For any  $j$ , let  $\{\mathcal{C}_j^i\}_{i=1}^{I_j}$  be the connected components of  $\{f_j \geq \lambda_* + \epsilon\}$  and  $\text{dist}(\mathcal{C}_j^i, \mathcal{C}_j^{i'}) \geq \sigma$  for  $i \neq i'$ .*

There exists subsets  $A_j$  and  $A'_j$  for density  $f_j$  such that  $A_{j,\sigma} \subset \mathcal{C}_j^i$  and  $A'_{j,\sigma} \subset \mathcal{C}_j^{i'}$  with the following additional property.

Consider any algorithm that is given  $n \geq 100$  i.i.d. samples  $\{X_i\}_{i=1}^n$  from some  $f_j \in F$  and, with probability at least  $3/4$ , outputs a tree in which the smallest cluster containing  $A_j \cap \{X_i\}_{i=1}^n$  is disjoint from the smallest cluster containing  $A'_j \cap \{X_i\}_{i=1}^n$ . Then there exists a constant  $C(d)$  only depending on  $d$  such that

$$n \geq \frac{C(d)}{\sigma^d \lambda^* \epsilon^2} \log \frac{1}{\sigma^d \lambda^*}. \quad (29)$$

The proof of the proposition is a straightforward modification of the proof of Theorem VI.1 of Chaudhuri et al. [2014] and we omit it for brevity. As a consequence, we conclude that DBSCAN achieves the optimal minimax scaling in both parameters  $\epsilon$  and  $\sigma$ .

## 5 Discussion

In this article we propose a new notion of consistency for estimating the clustering structure under various conditions. Our analysis shows that both the DBSCAN algorithm and the plug-in KDE cluster tree estimator is minimax optimal. Interestingly, the rates match, up to log terms, minimax rates for density estimation in the supreme norm for Hölder smooth densities. In particular, our results provide a complete, rigorous justification to the plausible belief, commonly held in density-based clustering, that clustering is as difficult as density estimation in the supreme norm.

In the rest of the discussion section, we will compare of our notion of separation with other existing ones in the literature.

### 5.1 $(\epsilon, \sigma)$ -separation in Chaudhuri et al. [2014]

The criterion of  $\delta$ -separation is most useful in the study of smooth densities. Nonetheless, it will be helpful to compare this with the notion of  $(\epsilon, \sigma)$ -separation defined in Chaudhuri et al. [2014], which is applicable to arbitrary densities.

**Definition 11** ( $(\epsilon, \sigma)$ -separation criterion in Chaudhuri et al. [2014]).

1. Let  $f$  be a density supported on  $X \subset \mathbb{R}^d$ . We say that  $A, A' \subset X$  are  $(\epsilon, \sigma)$ -separated if there exists  $S \subset X$  (the separator set) such that (i) any path in  $X$  from  $A$  to  $A'$  intersects  $S$ , and (ii)  $\sup_{x \in S_\sigma} f(x) < (1 - \epsilon) \inf_{x \in A_\sigma \cup A'_\sigma} f(x)$ .
2. Suppose an i.i.d samples  $\{X_i\}_{i=1}^n$  is given. An estimate of the cluster tree is said to be  $(\epsilon, \sigma)$  consistent if for any pair  $A$  and  $A'$  being  $(\epsilon, \sigma)$  separated, the smallest cluster containing  $A \cap \{X_i\}_{i=1}^n$  is disjoint from the smallest cluster containing  $A' \cap \{X_i\}_{i=1}^n$ .

In the following lemma we make a straightforward connection between the  $\delta$ -separation and  $(\epsilon, \sigma)$ -separation.

**Lemma 15.** Assume that  $p \in \Sigma(L, \alpha)$  with  $\alpha \leq 1$  and that  $A$  and  $A'$  are  $\delta$ -separated. Then,  $A$  and  $A'$  are  $(\epsilon, \sigma)$ -separated with

$$S = \{x: p(x) \leq \lambda - \delta\}, \quad \epsilon = \delta/(3\lambda) \quad \text{and} \quad \sigma^\alpha = \delta/(3L), \quad (30)$$

where  $\lambda = \inf_{z \in A \cup A'} p(z)$ .

*Proof of lemma 15.* Denote  $\lambda = \inf_{z \in A \cup A'} p(z)$ . Suppose for the sake of contradiction that there is a path  $l$  connects  $A$  and  $A'$  and that  $l \cap \{p \leq \lambda - \delta\} = \emptyset$ . Then by the continuity of  $p$  and the compactness of  $l$ , there exist  $\gamma > 0$  such that  $l \subset \{p \geq \lambda - \delta + \gamma\}$ . Thus  $A$  and  $A'$  belongs to the same path connected component of  $\{p \geq \lambda - \delta + \gamma\}$ . Since  $\{p \geq \lambda - \delta + \gamma\} \subset \{p > \lambda - \delta\}$ ,  $A$  and  $A'$  belongs to the same path connected component of  $\{p > \lambda - \delta\}$ . Since  $\{p > \lambda - \delta\}$  is an open set,  $A$  and  $A'$  be belongs to the same connected component of  $\{p > \lambda - \delta\}$ . This is a contradiction.

Let  $\sigma^\alpha = \delta/(3L)$  and  $\epsilon = \delta/3$ , then for any  $x \in S_\sigma$ ,  $p(x) \leq \lambda - \delta + L\sigma^\alpha = \lambda - 2\delta/3$ . Similarly if  $x \in A_\sigma \cup A'_\sigma$ ,  $p(x) \geq \lambda - L\sigma^\alpha = \lambda - \delta/3$ . Thus

$$(1 - \epsilon) \inf_{x \in A_\sigma \cup A'_\sigma} f(x) > (1 - \epsilon)\lambda - \delta/3 = \lambda - 2\delta/3 > \sup_{x \in S_\sigma} f(x)$$

□

According to the separation criterion in Definition 11, for a given  $\delta$ , two clusters can be  $\delta$ -separated for many values of  $\epsilon$  and  $\sigma$ . In particular, by taking the separator set to be large, it is easy to produce examples of  $\delta$ -separated clusters that are also  $(\epsilon, \sigma)$ -separated such that  $\delta$  is big but  $\sigma$  is small. This is simply because  $\sigma$  is heavily associated with  $S$ . And conversely, by taking an almost flat density function, it is possible to have a very large  $\sigma$  and very small  $\delta$ .

We remark that when  $\alpha > 1$ , there is no obvious relationship between the parameter  $(\sigma, \epsilon)$  in Definition 11 and  $\delta$  in Definition 6 as that in Lemma 15. For  $\alpha > 1$ , while  $p \in \Sigma(L, \alpha)$  implies that  $p$  is Lipschitz continuous, the Lipschitz constant in this case does not depend on  $L$  and  $\alpha$  in a simple manner. As a result, the parameter  $\sigma$ , representing the distance between connected components of upper level sets of  $p$ , is not related to  $\delta$  in any straightforward way.

## 5.2 Merge height distance

The notion of  $\delta$ -separation is closely related to the merge height first introduced by Eldridge et al. [2015] (see also Kim et al. [2016]). In the context of hierarchical clustering, merge height is used to described the height at which two points or two clusters merge into one cluster.

**Definition 12.** Let  $p$  be any density and  $T_p(\lambda)$  denote the cluster tree generated by  $p$  at level  $\lambda$ . For any two clusters  $A, A' \in T_p$ , their merge height  $m_p(A, A')$  is defined as

$$m_f(A, A') = \sup\{\lambda \in \mathbb{R} : \text{there exists } C \in T_p(\lambda) \text{ such that } A, A' \subset C.\}$$

It is easy to see that if two clusters of  $p$ ,  $A$  and  $A'$ , are  $\delta$ -separated, then their merge height is at most  $\lambda - \delta$ , where  $\lambda = \inf_{x \in A \cup A'} p(x)$ .

## References

- Ery Arias-Castro, David Mason, and Bruno Pelletier. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *The Journal of Machine Learning Research*, 17(1):1487–1514, 2016.
- ÍLLO Ba, Antonio Cuevas, Ana Justel, et al. Set estimation and nonparametric detection. *Canadian Journal of Statistics*, 28(4):765–782, 2000.



- Sivaraman Balakrishnan, Srivatsan Narayanan, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. Cluster trees on manifolds. In *Advances in Neural Information Processing Systems*, 2012.
- José E Chacón et al. A population background for nonparametric density-based clustering. *Statistical Science*, 30(4):518–532, 2015.
- Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada.*, pages 343–351, 2010.
- Kamalika Chaudhuri, Sanjoy Dasgupta, Samory Kpotufe, and Ulrike von Luxburg. Consistent procedures for cluster tree estimation and pruning. *IEEE Transactions on Information Theory*, 60(12):7900–7912, 2014.
- Yen-Chi Chen, Christopher R Genovese, and Larry Wasserman. Density level sets: Asymptotics, inference, and visualization. *Journal of the American Statistical Association*, (just-accepted), 2016.
- Antonio Cuevas. Set estimation: Another bridge between statistics and geometry. *Boletín de Estadística e Investigación Operativa*, 25(2):71–85, 2009.
- Antonio Cuevas and Ricardo Fraiman. A plug-in approach to support estimation. *The Annals of Statistics*, pages 2300–2312, 1997.
- Antonio Cuevas and Alberto Rodríguez-Casal. On boundary estimation. *Advances in Applied Probability*, 36(2):340–354, 2004.
- Antonio Cuevas, Wenceslao González-Manteiga, and Alberto Rodríguez-Casal. Plug-in estimation of general level sets. *Australian & New Zealand Journal of Statistics*, 48(1):7–19, 2006.
- Luc Devroye and Gary L. Wise. Detection of abnormal behavior via nonparametric estimation of the support. *SIAM Journal on Applied Mathematics*, 38(3):480–488, 1980.
- Manfredo Perdigao Do Carmo. *Riemannian geometry*. Birkhauser, 1992.
- Justin Eldridge, Mikhail Belkin, and Yusu Wang. Beyond hartigan consistency: Merge distortion metric for hierarchical clustering. In *Proceedings of The 28th Conference on Learning Theory*, pages 588–606, 2015.
- Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Kdd, pages 226–231, 1996.
- Junhao Gan and Yufei Tao. Dbscan revisited: Mis-claim, un-fixability, and approximation. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, pages 519–530, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-2758-9. doi: 10.1145/2723372.2737792. URL <http://doi.acm.org/10.1145/2723372.2737792>.
- Evarist Giné and Armelle Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'IHP Probabilités et statistiques*, 38:907–921, 2002.

- John A Hartigan. Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association*, 76(374):388–394, 1981.
- Heinrich Jiang. Density level set estimation on manifolds with dbscan. *arXiv preprint arXiv:1703.03503*, 2017a.
- Heinrich Jiang. Uniform convergence rates for kernel density estimation. In *International Conference on Machine Learning*, pages 1694–1703, 2017b.
- Jisu Kim, Yen-Chi Chen, Sivaraman Balakrishnan, Alessandro Rinaldo, and Larry Wasserman. Statistical inference for cluster trees. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1839–1847. Curran Associates, Inc., 2016.
- Jussi Klemelä. Complexity penalized support estimation. *Journal of multivariate analysis*, 88(2):274–297, 2004.
- Jussi Klemelä. *Smoothing of Multivariate Data: Density Estimation and Visualization*. Wiley, 2009.
- Aleksandr Petrovich Korostelev and Alexandre B Tsybakov. *Minimax theory of image reconstruction*, volume 82. Springer Science & Business Media, 1993.
- Alexander Korostelev and Michael Nussbaum. The asymptotic minimax constant for sup-norm loss in nonparametric density estimation. *Bernoulli*, 5(6):1099–1118, 1999.
- Samory Kpotufe and Ulrike V Luxburg. Pruning nearest neighbor cluster trees. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 225–232, 2011.
- Yukio Matsumoto. *An introduction to Morse theory*, volume 208. American Mathematical Soc., 2002.
- James R Munkres. *Topology*. Prentice Hall, 2000.
- Laurent Najman and Michel Couprie. Building the component tree in quasi-linear time. *IEEE Transactions on image processing*, 15(11):3531–3539, 2006.
- Deborah Nolan and David Pollard. U-processes: rates of convergence. *The Annals of Statistics*, pages 780–799, 1987.
- Mathew D. Penrose. Single linkage clustering and continuum percolation. *Journal of Multivariate Analysis*, 53:94–109, 1995.
- Wolfgang Polonik. Measuring mass concentrations and estimating density contour clusters-an excess mass approach. *The Annals of Statistics*, pages 855–881, 1995.
- Philippe Rigollet and Régis Vert. Optimal rates for plug-in estimators of density level sets. *Bernoulli*, pages 1154–1178, 2009.
- Alessandro Rinaldo and Larry Wasserman. Generalized density clustering. *The Annals of Statistics*, pages 2678–2722, 2010.

- Alessandro Rinaldo, Aarti Singh, Rebecca Nugent, and Larry Wasserman. Stability of density-based clustering. *The Journal of Machine Learning Research*, 13(1):905–948, 2012.
- Artri Singh, Clayton Scott, Robert Nowak, and Aarti Singh. Adaptive hausdorff estimation of density level sets. *The Annals of Statistics*, 37(5B):2760–2782, 2009.
- Bharath K Sriperumbudur and Ingo Steinwart. Consistency and rates for clustering with dbscan. In *AISTATS*, pages 1090–1098, 2012.
- Werner Stuetzle. Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *Journal of classification*, 20(1):025–047, 2003.
- Werner Stuetzle and Rebecca Nugent. A generalized single linkage method for estimating the cluster tree of a density. *Journal of Computational and Graphical Statistics*, 19(2), 2010.
- Alexandre B Tsybakov. Introduction to nonparametric estimation. revised and extended from the 2004 french original. translated by vladimir zaiats, 2009.
- Alexandre B Tsybakov et al. On nonparametric estimation of density level sets. *The Annals of Statistics*, 25(3):948–969, 1997.
- Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of Complexity*, pages 11–30. Springer, 2015.
- Bingchen Wang, Chenglong Zhang, Lei Song, Zhao Lianhe, Yu Dou, and Zihao Yu. Design and optimization of dbscan algorithm based on cuda. 06 2015. arXiv preprint arXiv:1506.02226.
- RM Willett and Robert D Nowak. Minimax optimal level-set estimation. *IEEE Transactions on Image Processing*, 16(12):2965–2979, 2007.
- Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

## A Topological Preliminaries

For completeness, we review the definition of connectedness from the general topology.

**Definition 13** (Munkres [2000] Chapter 3). *Let  $U$  be any nonempty subset in  $\mathbb{R}^d$ . Then  $U$  is said to be connected, if, for every pair of open subsets  $A, A'$  of  $U$  such that  $A \cup A' = U$ , we have either  $A = \emptyset$  or  $A' = \emptyset$ . The maximal connected subsets of  $U$  are called the connected components of  $U$ .*

We briefly explain why the connected components naturally introduce a hierarchical structure to the level sets of  $p$ . Let  $\lambda_1 > \lambda_2$ , so we have  $\{p \geq \lambda_1\} \subset \{p \geq \lambda_2\}$ .

- Suppose  $A$  is any subset of  $\mathbb{R}^d$ , and  $A$  belongs to the same connected component of  $\{p \geq \lambda_1\}$ . Then  $A$  is contained in the same connected component of  $\{p \geq \lambda_2\}$ .
- Suppose  $A \cup A' \subset \{p \geq \lambda_1\}$  and they belong to distinct connected components of  $\{p \geq \lambda_2\}$ . Then  $A$  and  $A'$  are not contained in the same connected component of  $\{p \geq \lambda_1\}$ .

We also review a closed related concepts, which is call the path connectedness in general topology.

**Definition 14.** We say that a subset  $U \subset \mathbb{R}^d$  is path connected if for any  $x, y \in U$ , there exists a path continuous  $\mathcal{P} : [0, 1] \rightarrow U$  such that  $\mathcal{P}(0) = x$  and  $\mathcal{P}(1) = y$ .

The main reason we introduce the path connectedness is that if  $U$  is an open set in  $\mathbb{R}^d$ , then  $U$  is connected if and only if it is path connected. Therefore a simple but useful consequence is that for any  $\lambda$ , the connected components of  $\{p > \lambda\}$  are also the path connected components. We will repeatedly use these topological properties in our analysis without further mentioning. The proofs of them are omitted and can be found in [Munkres \[2000\]](#) or any other books on general topology.

## B Proofs in Sections 2 and 3

*Proof of Lemma 1.* All the claims of the lemma follow from the simple observation that  $\widehat{L}(\lambda_k)$  is a union of collection of balls of radius  $h$  and centered at  $\widehat{D}(\lambda_k)$ . Thus  $X_i$  and  $X_j$  are in the same connected component of  $\widehat{L}(\lambda_k)$  if and only if there exists  $\{X_{k_1}, \dots, X_{k_L}\}$  such that  $X_{k_1} = X_i, X_{k_L} = X_j$  and that  $\|X_{k_l} - X_{k_{l+1}}\| \leq 2h$  for all  $1 \leq l \leq L - 1$ .  $\square$

### B.1 Proofs from Section 3.1

We begin by justifying the event [Proposition 16](#).

**Proposition 16.** Let  $K$  be the spherical kernel,  $h > 0$  a fixed bandwidth and  $\hat{p}_h$  be the corresponding kernel density estimator. Then

$$\mathbb{P} \left( \sup_{x \in \mathbb{R}^d} |\hat{p}_h(x) - p_h(x)| \leq a_n \right) \geq 1 - \gamma, \quad (31)$$

where  $a_n$  is defined in [\(9\)](#).  $\|p\|_\infty = \sup_{x \in \mathbb{R}^d} p(x)$ .

**Remark 6.** The quantity  $\|p\|_\infty$  can be replaced by the smaller quantity  $\sup_{x \in \mathbb{R}^d} p_h(x)$ , for any  $h > 0$ .

*Proof of Proposition 16.* Let  $\mathcal{B}$  denote the collection of all closed Euclidean balls in  $\mathbb{R}^d$ . Then, the relative VC bounds for balls (see [Vapnik and Chervonenkis \[2015\]](#)) in  $\mathbb{R}^d$  states that, for any  $\gamma \in (0, 1)$ ,

$$\sup_{B \in \mathcal{B}} \frac{P_n(B) - P(B)}{\sqrt{P_n(B)}} \leq 2\sqrt{\frac{(d+1)\log(2n+1) + \log(8/\gamma)}{n}}$$

and

$$\sup_{B \in \mathcal{B}} \frac{P(B) - P_n(B)}{\sqrt{P(B)}} \leq 2\sqrt{\frac{(d+1)\log(2n+1) + \log(8/\gamma)}{n}},$$

with probability  $1 - \gamma$ . Since  $\hat{p}_h(x) = \frac{1}{nh^d V_d} \sum_{i=1}^n \mathbf{1}_{\{X_i \in B(x, h)\}} = \frac{1}{h^d V_d} P_n(B(x, h))$  and  $p_h(x) = E(\hat{p}_h(x)) = \frac{1}{h^d V_d} P(B(x, h))$ , it follows that, with probability at least  $1 - \gamma$ ,

$$\sup_{x \in \mathbb{R}^d} \frac{\hat{p}_h(x) - p_h(x)}{\sqrt{\hat{p}_h(x)}} \leq 2\sqrt{\frac{(d+1)\log(2n+1) + \log(8/\gamma)}{nh^d V_d}}$$

and

$$\sup_{x \in \mathbb{R}^d} \frac{p_h(x) - \hat{p}_h(x)}{\sqrt{p_h(x)}} \leq 2\sqrt{\frac{(d+1)\log(2n+1) + \log(8/\gamma)}{nh^d V_d}}.$$

Since  $p_h(x) \leq p_{\max}$  for all  $x \in \mathbb{R}^d$ ,

$$\|\hat{p}_h - p_h\|_\infty \leq 2\sqrt{(p_{\max} + \|\hat{p}_h - p_h\|_\infty) \frac{(d+1)\log(2n+1) + \log(8/\gamma)}{nh^d V_d}} \quad (32)$$

Solving this quadratic inequality in  $\|\hat{p}_h - p_h\|_\infty$  gives

$$\|\hat{p}_h - p_h\|_\infty \leq 4\frac{(d+1)\log(2n+1) + \log(8/\gamma)}{nh^d V_d} + 2\sqrt{p_{\max} \frac{(d+1)\log(2n+1) + \log(8/\gamma)}{nh^d V_d}},$$

as claimed.  $\square$

We begin by justifying (7). Since this is a well known result, we simply use a result of [Sriperumbudur and Steinwart \[2012\]](#). We will assume the following condition for the kernel  $K$  which is fairly standard in the non-parametric literature.

VC. The kernel  $K : \mathbb{R}^d \rightarrow \mathbb{R}$  has bounded support and integrates to 1. Let  $\mathcal{F}$  be the class of functions of the form

$$z \in \mathbb{R}^d \mapsto K(x - z), \quad z \in \mathbb{R}^d.$$

Then,  $\mathcal{F}$  is a uniformly bounded VC class: there exist positive constants  $A$  and  $v$  such that

$$\sup_P \mathcal{N}(\mathcal{F}, L^2(P), \epsilon \|F\|_{L^2(P)}) \leq (A/\epsilon)^v,$$

where  $\mathcal{N}(T, d, \epsilon)$  denotes the  $\epsilon$ -covering number of the metric space  $(T, d)$ ,  $F$  is the envelope function of  $\mathcal{F}$  and the sup is taken over the set of all probability measures on  $\mathbb{R}^d$ . The constants  $A$  and  $v$  are called the VC characteristics of the kernel.

The assumption VC holds for a large class of kernels, including any compact supported polynomial kernel and the Gaussian kernel. See [Nolan and Pollard \[1987\]](#) and [Giné and Guillou \[2002\]](#).

**Proposition 17** ([Sriperumbudur and Steinwart \[2012\]](#)). *Let  $P$  be the probability measure on  $\mathbb{R}^d$  with Lebesgue density bounded by  $\|p\|_\infty$  and assume that the kernel  $K$  belongs to  $L^\infty(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$  satisfies the VC assumption. Then for any  $\gamma > 0$  and  $h > 0$ , there exists an absolute constant  $C$  depending on the VC characteristic of  $K$  such that, with probability no smaller than  $1 - e^{-\gamma}$ ,*

$$\|p_h - \hat{p}_h\|_\infty \leq \frac{C}{nh^d} \left( \gamma + v \log \frac{2A}{\sqrt{h^d \|p\|_\infty \|K\|_2^2}} \right) + C \sqrt{\frac{2\|p\|_\infty}{nh^d}} \left( \gamma \|K\|_\infty^2 + v \|K\|_2^2 \log \frac{2A}{\sqrt{h^d \|p\|_\infty \|K\|_2^2}} \right)$$

## B.2 Proofs from Section 3.3

*Proof of Proposition 2.* To show Proposition 2, we first establish simple topological facts.

**Lemma 18.** *Suppose  $p : \mathbb{R}^d \rightarrow \mathbb{R}$  are compactly supported. If  $A$  and  $A'$  are in the same connected components of  $\{p \geq \lambda_i\}$  for  $i = 1, 2, \dots, \infty$  and that  $\lambda_i \leq \lambda_{i+1}$ , then  $A$  and  $A'$  are in the same connected components of  $\{p \geq \lambda_0\}$ , where  $\lambda_0 = \sup_i \lambda_i$ .*

*Proof of lemma 18.* Let  $\mathcal{C}_i$  be the connected component of  $\{p \geq \lambda_i\}$  that contains  $A$  and  $A'$ . Thus  $\mathcal{C}_i$  are compact and connected. Since  $\mathcal{C}_{i+1} \subset \mathcal{C}_i$  for all  $i \geq 1$ ,  $\bigcap_{i=1}^{\infty} \mathcal{C}_i$  is connected. Thus  $A, A' \subset \bigcap_i \mathcal{C}_i \subset \{p \geq \lambda_0\}$ .  $\square$

Consider

$$\lambda^* = \sup\{\lambda : A, A' \text{ belongs to the same connected components of } \{p \geq \lambda\}\}$$

Then  $\lambda_2 \leq \lambda^* \leq \lambda_1$ . By lemma 18,  $A$  and  $A'$  are in the same connected components of  $\{p \geq \lambda^*\}$ . Thus  $\lambda_2 \leq \lambda^* < \lambda_1$ .

In order to show that  $\lambda^*$  is split level, it suffices to show that  $A$  and  $A'$  are in the different connected components of  $\{p > \lambda^*\}$ . Suppose for the sake of contradiction that  $A$  and  $A'$  are connected in  $\{p > \lambda^*\}$ . Then  $A$  and  $A'$  are path connected as  $\{p > \lambda^*\}$  is open. Thus there exist  $\mathcal{P}$  connects  $A$  and  $A'$  in  $\{p > \lambda^*\}$ . Since  $\mathcal{P}$  is compact,  $p(\mathcal{P}) > \lambda^*$  implies that there exists  $a > 0$  such that  $\lambda^* + a < \lambda_1$  and  $\mathcal{P}, A, A' \subset \{p \geq \lambda^* + a\}$ . Thus  $A$  and  $A'$  belong to the same connected component of  $\{p \geq \lambda^* + a\}$ . This is a contradiction because by construction of  $\lambda^*$ ,  $A$  and  $A'$  belongs to the different connected components of  $\{p \geq \lambda^* + a\}$ .  $\square$

*Proof of corollary 3.* Suppose  $A$  and  $A'$  are  $\delta$  separated with respect to  $\lambda$ . Then  $A$  and  $A'$  belongs to distinct connected components of  $\{p > \lambda - \delta\}$  where  $\lambda = \inf_{x \in A \cup A'} f(x)$ . Let  $0 < \epsilon \leq \delta$  be given. Then since  $\{p \geq \lambda - \delta + \epsilon\} \subset \{p > \lambda - \delta\}$ ,  $A$  and  $A'$  belongs to distinct connected components of  $\{p \geq \lambda - \delta + \epsilon\}$ .

Since  $\mathbb{R}^d = \{p \geq 0\}$  is connected,  $A$  and  $A'$  belongs to the same connected component of  $\{p \geq 0\}$ . By proposition 2, there exists  $0 \leq \lambda^* < \lambda - \delta + \epsilon$  such that  $A$  and  $A'$  in the same connected component of  $\{p \geq \lambda^*\}$  and in different connected components of  $\{p > \lambda^*\}$ . By taking  $\epsilon \rightarrow 0$ , the claimed result follows.  $\square$

### B.3 Proofs from sections 3.4

*Proof of Lemma 4.* From the proof of lemma 7, it can be see that

$$\left\{ p \geq \lambda + C \frac{\log(n)}{n^{\alpha/(2\alpha+d)}} \right\} \cap \{X_i\}_{i=1}^n \subset \hat{D}(\lambda) \subset \left\{ p \geq \lambda - C \frac{\log(n)}{n^{\alpha/(2\alpha+d)}} \right\}.$$

Thus for any  $y \in B(X_j, h)$  for some  $X_j \in \hat{D}(\lambda)$ ,

$$p(y) \geq p(X_j) - Lh^\alpha \geq \lambda - C \frac{\log(n)}{n^{\alpha/(2\alpha+d)}} - Lh^\alpha,$$

where the first inequality follows from  $|p(y) - p(X_j)| \leq Lh^\alpha$ . Therefore the above display implies

$$\bigcup_{X_j \in \hat{D}(\lambda)} B(X_j, h) \subset \left\{ p \geq \lambda - C \frac{\log(n)}{n^{\alpha/(2\alpha+d)}} - Lh^\alpha \right\}.$$

For the other inclusion, let  $x \in \left\{ p \geq \lambda + C \frac{\log(n)}{n^{\alpha/(2\alpha+d)}} + Lh^\alpha \right\}$ . Then  $\hat{p}_h(x) \geq \lambda + Lh^\alpha$ . Thus  $B(x, h) \cap \{X_i\}_{i=1}^n \neq \emptyset$ , or else  $\hat{p}_h(x) = 0$ . Let  $X_j \in B(x, h)$ . Therefore

$$p(X_j) \geq p(x) - Lh^\alpha \geq \lambda + \frac{\log(n)}{n^{\alpha/(2\alpha+d)}}.$$

Thus  $\hat{p}_h(X_j) \geq \lambda$ , which means that  $X_j \in \hat{D}(\lambda)$ . So  $x \in \bigcup_{X_j \in \hat{D}(\lambda)} B(X_j, h)$  and the first inclusion follows. □

*Proof of Theorem 6.* Let  $\mathcal{B}$  be the event that

$$\mathcal{B} = \left\{ \sup_{x \in \mathbb{R}^d} |\hat{p}_h(x) - p(x)| \leq a_n \right\}$$

We can choose  $a_n$  so that  $P(\mathcal{B}) \geq 1 - 1/n$  and that  $a_n = O(\log(n)n^{-\alpha/(d+2\alpha)})$ . All the argument will be made under the good event  $\mathcal{B}$ .

Observe that  $p$  has connected support. Therefore,  $\lambda = 0$  is not a split level. Assume that  $\lambda_0 = \min\{\lambda^* : \lambda^* \text{ is a split level of } p\}$ . Then  $\lambda_0 > 0$ . If  $h = O(n^{-1/(2\alpha+d)})$ , for large  $n$ , we have  $2a_n + (4h/c_S)^\alpha < \min\{\delta_S, \delta_0\}$ . Take

$$\delta \geq 2a_n + (4h/c_S)^\alpha/c_S.$$

Let  $A$  and  $A'$  are two sets being  $\delta$ -separated and let  $\lambda = \inf_{x \in A \cup A'} p(x)$ . Since  $A$  and  $A'$  are in distinct connected components of  $\{p > \lambda - \delta\}$ , by proposition 2 there exists  $\lambda^*$  being a split level of  $p$  such that  $\lambda^* \leq \lambda - \delta$  and that  $A$  and  $A'$  belongs to distinct connected components of  $\{p > \lambda^*\}$ . Thus  $A$  and  $A'$  belong to distinct connected components of  $\{p > \lambda'\}$ , where

$$\lambda' = \lambda^* + 2a_n + (4h/c_S)^\alpha/c_S.$$

Let  $\{\mathcal{C}_k\}_{k=1}^K$  be the collection of connected components of  $\{p > \lambda'\}$ . Thus we have  $A \subset \mathcal{C}_k$  and  $A' \subset \mathcal{C}_{k'}$  for some  $k \neq k'$ . In order to show the smallest cluster containing  $A \cap \{X_i\}_{i=1}^n$  and  $A' \cap \{X_i\}_{i=1}^n$  are disjoint with high probability, it suffices to show the following statement.

- Let  $A$  and  $A'$  be two connected subsets of  $\{p > \lambda'\}$  and belong to two distinct connected components of  $\{p > \lambda^*\}$ . Then the smallest cluster containing  $A \cap \{X_i\}_{i=1}^n$  and  $A' \cap \{X_i\}_{i=1}^n$  are disjoint with high probability.

Note that this observation reduce the original statement which concerns with generic  $\delta$ -separated sets to the current statement which only concerns with one level near the split level. Since there are finitely many split levels, a simple union bound will suffice to show the  $\delta$  consistency of the cluster tree returned by Algorithm 2.

The proof will be completed by the following two claims.

**Claim 1.** If  $A$  is a connected subset of  $\{p > \lambda'\}$ , then  $A \cap \{X_i\}_{i=1}^n$  is in the same connected component of

$$\hat{L}(\lambda' - a_n) := \bigcup_{\{X_j : \hat{D}(\lambda' - a_n)\}} B(X_j, 2h). \quad (33)$$

*Proof.* It suffices to show that

$$\{p > \lambda'\} \subset \widehat{L}(\lambda' - a_n). \quad (34)$$

Since for large  $n$ ,

$$a_n + (4h/c_S)^\alpha \leq \delta_0,$$

By **C2** there exists  $\mathcal{N}_h \subset \{p > \lambda'\}$  with  $\text{card}(\mathcal{N}_h) \leq A_c(h)^{-d}$  such that  $\mathcal{N}_h$  is a  $h$  cover. Since  $\{p > \lambda'\}$  satisfies the inner cone condition **C1**,

$$P(B(x, h) \cap \{p > \lambda'\}) \geq \lambda^* c_I V_d h^d \geq \lambda_0 c_I V_d h^d.$$

So there exists  $c'_I$  only depending on  $d$  and  $c_I$  such that

$$P(\{\{X_i\}_{i=1}^n \cap B(x, h) \cap \{p > \lambda'\} = \emptyset\}) \leq (1 - \lambda_0 c_I V_d h^d)^n \leq \exp(-c'_I \lambda_0 n^{2\alpha/(\alpha+d)}) = o(n^{-2}),$$

where the second inequality follows from  $h = O(n^{1/(2\alpha+d)})$  and the equality follows from  $\lambda_0 n^{2\alpha/(2\alpha+d)} / \log(n) \rightarrow \infty$  and  $n$  being large enough. Consider the event

$$\mathcal{A} = \{\{X_i\}_{i=1}^n \cap B(x, h) \cap \{p > \lambda'\} \neq \emptyset \text{ for all } x \in \mathcal{N}_h\}.$$

By the union bound

$$P(\mathcal{A}^c) \leq \text{card}(\mathcal{N}_h) \exp(-c'_I \lambda_0 n^{2\alpha/(2\alpha+d)}) = A_c h^{-d} \exp(-c'_I \lambda_0 n^{2\alpha/(2\alpha+d)}) = o(1). \quad (35)$$

So for any  $y \in \{p > \lambda'\}$ , there exists  $x \in \mathcal{N}_h$  such that  $|y - x| \leq h$ . Under event  $\mathcal{A}$  there exists  $X_j \in \{p > \lambda'\}$  such that  $|X_j - x| \leq h$ . Therefore  $y \in B(X_j, 2h)$ . Since

$$X_j \in \{X_i\}_{i=1}^n \cap \{p > \lambda'\} \subset \widehat{D}(\lambda' - a_n),$$

the claim follows.  $\square$

To finish the proof of the theorem, we still need to show at level  $\lambda' - a_n$  the data points  $A \cap \{X_i\}_{i=1}^n$  and  $A' \cap \{X_i\}_{i=1}^n$  are contained in distinct clusters. Therefore the following claim finish the proof.

**Claim 2.** There exists a partition  $\{S_i\}_{i=1}^I$  of  $\widehat{D}(\lambda' - a_n)$  such that  $A \cap \{X_i\}_{i=1}^n$  and  $A' \cap \{X_i\}_{i=1}^n$  belong to distinct subsets of the partition and that data points in distinct subsets of the partition are mutually disconnected.

*Proof.* Let  $\{B_i\}_{i=1}^I$  be the collection of connected components of  $\{p \geq (4h/c_S)^\alpha + \lambda^*\}$ . Since  $A$  and  $A'$  belong to distinct connected components of  $\{p > \lambda^*\}$ , and  $\lambda^* < (4h/c_S)^\alpha + \lambda^*$ ,  $A$  and  $A'$  are contained in distinct elements of  $\{B_i\}_{i=1}^I$ . From condition **S**,

$$\min_{i \neq j} \text{dist}(B_i, B_j) \geq 4h. \quad (36)$$

Note that  $\widehat{D}(\lambda' - a_n) \subset \{p \geq (4h/c_S)^\alpha + \lambda^*\}$  as a consequence of event  $\mathcal{B}$ . Thus  $S_i = B_i \cap \widehat{D}(\lambda' - a_n)$  form a partition of  $\widehat{D}(\lambda' - a_n)$ . Let

$$L_i = \bigcup_{X_j \in S_i} B(X_j, 2h).$$

By (36),  $L_i \cap L_j = \emptyset$  if  $i \neq j$ . This shows that data points in distinct subsets of the partition  $\{S_i\}_{i=1}^I$  are mutually disconnected at the graph  $\mathbb{C}(h, \lambda' - a_n)$ .  $\square$



□

**Proposition 19.** *Suppose  $p$  is a Morse function, then  $p$  satisfies **C** and **S(2)**.*

*Proof of Proposition 19.*

**Step 1.** In this step we show that condition **S(2)** holds. Consider an arbitrary split level  $\lambda$ , and two connected components  $C_1, C_2$ . If

$$\inf_{\delta > 0} \text{dist}(C_1 \cap \{p \geq \lambda + \delta\}, C_2 \cap \{p \geq \lambda + \delta\}) > 0$$

then we have  $\text{dist}(C_1 \cap \{p \geq \lambda\}, C_2 \cap \{p \geq \lambda\}) > 0$ , and the thesis is trivial. Thus assume that

$$\inf_{\delta > 0} \text{dist}(C_1 \cap \{p \geq \lambda + \delta\}, C_2 \cap \{p \geq \lambda + \delta\}) = 0,$$

i.e.

$$\lim_{\delta \rightarrow 0} \text{dist}(C_1 \cap \{p \geq \lambda + \delta\}, C_2 \cap \{p \geq \lambda + \delta\}) = 0.$$

Thus there exists  $y_0 \in \{p = \lambda\}$ , and points  $y_{1,2}^\delta \in C_{1,2} \cap \{p \geq \lambda + \delta\}$  such that

$$y_{1,2}^\delta \xrightarrow{\delta \rightarrow 0} y_0, \quad |y_1^\delta - y_2^\delta| = \text{dist}(C_1 \cap \{p \geq \lambda + \delta\}, C_2 \cap \{p \geq \lambda + \delta\}).$$

It is straightforward to check that  $p(y_1^\delta) = p(y_2^\delta) = \lambda + \delta$ .

The thesis is now rewritten as  $|y_1^\delta - y_2^\delta| \geq c_S \delta^{1/2}$  for some constant  $c_S > 0$  and all sufficiently small  $\delta$ . Since split levels are also critical,  $\nabla p(y_0) = 0$ ; since  $p$  is a Morse function,  $\nabla^2 p(y_0)$  is non-degenerate. By Taylor formula we have

$$\delta = p(y_j^\delta) - p(y_0) = (y_j^\delta - y_0)^T \nabla^2 p(y_0) (y_j^\delta - y_0) / 2 + O(|y_j^\delta - y_0|^3), \quad j = 1, 2, \quad (37)$$

and, as  $\nabla^2 p(y_0)$  is non-degenerate, it follows  $|y_j^\delta - y_0| = O(\delta^{1/2})$ , i.e. there exist constants  $c_1, c_2, \delta_0 > 0$  such that

$$c_1 \delta^{1/2} \leq |y_j^\delta - y_0| \leq c_2 \delta^{1/2} \quad \text{for all } \delta \in (0, \delta_0).$$

We can estimate  $c_2$  from below: denoting by

$$a := \max\{|e_1(y_0)|, |e_2(y_0)|\}, \quad e_1(y_0), e_2(y_0) = \text{eigenvalues of } \nabla^2 p(y_0),$$

(37) gives

$$(y_j^\delta - y_0)^T \nabla^2 p(y_0) (y_j^\delta - y_0) \leq a c_2^2 |y_j^\delta - y_0|^2,$$

hence  $c_2 \geq \sqrt{2/a}$ . By the Lipschitz regularity of the gradient, i.e. hypothesis

$$|\nabla p(x) - \nabla p(y)| \leq L|x - y|$$

for some  $L > 0$ , we have

$$|\nabla p(y_1^\delta) - \nabla p(y_0)| = |\nabla p(y_1^\delta)| \leq L|y_1^\delta - y_0| \leq Lc_2 \delta^{1/2}. \quad (38)$$

Consider now the segment  $[y_1^\delta, y_2^\delta]$  between  $y_1^\delta$  and  $y_2^\delta$ : since  $y_j^\delta \in C_j \cap \{p \geq \lambda + \delta\}$  ( $j = 1, 2$ ), and  $C_j \cap \{p \geq \lambda + \delta\}$  are disconnected for all  $\delta > 0$ , there exists some point  $z \in [y_1^\delta, y_2^\delta]$  such that  $p(z) < \lambda + \delta/2$ . By Taylor's formula we then have

$$p(z) = p(y_1^\delta) + \nabla p(y_1^\delta) \cdot (z - y_1^\delta) + (z - y_1^\delta)^T \nabla^2 p(y_1^\delta) (z - y_1^\delta) / 2 + O(|z - y_1^\delta|^3)$$

If inequality  $|y_1^\delta - y_2^\delta| \leq k\delta^{1/2}$  were to hold for some  $k > 0$ , then since the domain is compact and  $\nabla^2 p \in C^2$ , denoting by

$$A := \sup_x \left( \max\{|e_1(x)|, |e_2(x)|\} \right), \quad e_1(x), e_2(x) = \text{eigenvalues of } \nabla^2 p(x),$$

we have

$$\begin{aligned} |p(z) - p(y_1^\delta)| &\leq |\nabla p(y_1^\delta)| \cdot |z - y_1^\delta| + |\nabla^2 p(y_1^\delta)| \cdot |z - y_1^\delta|^2/2 \\ &\leq |\nabla p(y_1^\delta)| \cdot |y_1^\delta - y_2^\delta| + |\nabla^2 p(y_1^\delta)| \cdot |y_1^\delta - y_2^\delta|^2/2 \stackrel{(38)}{\leq} (Lkc_2 + k^2 A/2)\delta. \end{aligned}$$

Since  $p(y_1^\delta) = \lambda + \delta$ , and  $p(z) < \lambda + \delta/2$ , we need  $Lkc_2 + k^2 A/2 > 1/2$ , hence

$$k \geq A^{-1}(\sqrt{L^2 c_2^2 + A} - Lc_2),$$

i.e.

$$\begin{aligned} |y_1^\delta - y_2^\delta| &= \text{dist}(C_1 \cap \{p \geq \lambda + \delta\}, C_2 \cap \{p \geq \lambda + \delta\}) \\ &\geq A^{-1}(\sqrt{L^2 c_2^2 + A} - Lc_2)\delta^{1/2} \geq A^{-1}(\sqrt{2L^2/a + A} - L\sqrt{2/a})\delta^{1/2}. \end{aligned}$$

**Step 2.** In this step we show that condition **C** holds.

**Proof of C1.** Since a Morse function has only isolated non degenerate critical points, and an isolated set in a compact domain is also finite, we infer that  $\nabla p(x) = 0$  only for finitely many  $x$ . In particular, since  $\lambda^*$  are split levels, and  $\{p = \lambda^*\}$  contains a critical point, there exist sufficiently small  $\delta_1, \delta_2 > 0$  such that  $\{\lambda^* + \delta_1 \leq p \leq \lambda^* + \delta_2\}$  contains no critical points (since there are only finitely many critical points). Since the level sets are orthogonal to the gradient, we infer that  $\{p = \lambda^* + \delta_1\}$  is smooth. In particular,  $\{p = \lambda^* + \delta_1\}$  it satisfies the inner cone property with  $c_I = 1/2$ .

The key difficulty in extending the above argument to  $\{p > \lambda^*\}$  (instead of just  $\{p \geq \lambda^* + \delta_1\}$  with  $\delta_1 > 0$ ) is that the norm of gradient  $|\nabla p|$  can approach zero as  $\delta_1 \rightarrow 0$ , since  $\{p = \lambda^*\}$  is a split level, hence it contains critical points.

The Morse function requirement, however, gives the ‘‘bare minimum’’ regularity to ensure C1. We aim to prove, by contradiction, that  $\{p \geq \lambda^*\}$  also satisfies C1, i.e. the boundary  $\{p = \lambda^*\}$  does not exhibit cusps. If a cusp were to appear, then there exist arc-length parameterized curves  $\gamma_j : [0, \epsilon] \rightarrow \Omega$ ,  $j = 1, 2$ , such that  $x_0 = \gamma_1(0) = \gamma_2(0)$  and the angle  $\angle \gamma_1(s)x_0\gamma_2(s) \rightarrow 0$  as  $s \rightarrow 0$ .

Consider a level set  $\{p = \lambda^* + \delta\}$ , for small  $\delta > 0$ . Let  $y_\delta \in \{p = \lambda^* + \delta\}$  be the point on  $\{p = \lambda^* + \delta\}$  closest to  $x_0$ , i.e.  $|x_0 - y_\delta| = \min_{y \in \{p = \lambda^* + \delta\}} |x_0 - y|$ , and we proved that  $|x_0 - y_\delta| = O(\sqrt{\delta})$ . Let  $p_\delta, q_\delta$  be the intersection between  $\{p = \lambda^*\}$  and the tangent line to  $\{p = \lambda^* + \delta\}$  through  $y_\delta$ . Clearly, as  $\{p = \lambda^*\}$  has a cusp at  $x_0$ , we get  $\lim_{\delta \rightarrow 0} \angle p_\delta x_0 q_\delta = 0$ . Thus

$$\text{dist}(y_\delta, \{p = \lambda^*\}) \leq |p_\delta - y_\delta| = o(\sqrt{\delta}).$$

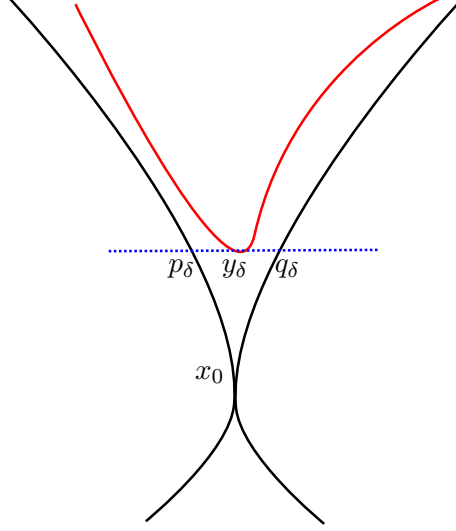


Figure 3: Construction in the proof of Morse function case.

Since  $\nabla p(x_0) = 0$ , and the gradient  $\nabla p$  is  $L$ -Lipschitz continuous for some constant  $L$ , we infer  $|\nabla p(y)| \leq A\sqrt{\delta}$  for some  $A > 0$  and all  $y$  on the segment  $[p_\delta, y_\delta]$ . Thus it follows

$$\delta = |p(p_\delta) - p(y_\delta)| \leq AL\sqrt{\delta}|p_\delta - y_\delta| = o(\delta).$$

This is a contradiction.

**Proof of C2.** Let  $U = \{p \geq \lambda^* + \delta\}$ . Fix an arbitrary  $r$ . Clearly  $U \subseteq \bigcup_{x \in U} B(x, r/3)$ . Since  $U = \{p \geq \lambda^* + \delta\}$  is closed, and the domain  $\Omega$  is compact, we infer  $U = \{p \geq \lambda^* + \delta\}$  is also compact. Thus we can extract a covering  $U \subseteq \bigcup_{i=1}^{C'_r} B(x_i, r/3)$  with finitely many balls. By Vitali covering lemma, we can further extract mutually disjoint balls  $B(x_{i_j}, r/3)$  such that

$$U \subseteq \bigcup_{j=1}^{C'_r} B(x_{i_j}, r)$$

Since  $B(x_{i_j}, r/3) \subset \Omega_{r/3}$ , and  $\{B(x_{i_j}, r/3)\}_{j=1}^{C'_r}$  are pairwise disjoint, we have

$$V_d C'_r (r/3)^{-d} \leq \mathcal{L}^d(\Omega_{r/3}).$$

Thus we can choose  $\mathcal{N}_r = \{x_{i_j}\}$ ,  $j = 1, \dots, C'_r$ . □

### B.3.1 Proofs in Section 3.6

*Proof of lemma 9.* Let  $\lambda > 0$  be given. Later in the proof, it can be seen that  $\lambda = C_0$ , being the common upper bound of  $f_i \in F$ . Define  $a > 0$  to be such that

$$56\lambda \cdot 8^{d-1} a^d = 1. \tag{39}$$

Consider

$$f(x) = \begin{cases} \lambda, & x \in [0, 56a] \times [0, 8a]^{d-1} = \Omega \\ 0, & \text{otherwise.} \end{cases} \quad (40)$$

Let  $b = \left(\frac{\log(32)}{n\lambda V_d}\right)^{1/d}$ . For  $0 < \alpha < 1$ , define

$$g(r) = \begin{cases} 0, & 0 \leq r \leq b \\ \mathcal{K} \left( \left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha} - |r - b - \left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}| \right)^\alpha, & |r - b - \left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}| \leq \left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha} \\ 0, & \text{otherwise,} \end{cases}$$

and for  $\alpha \geq 1$ , define

$$g(r) = \begin{cases} 0, & 0 \leq r \leq b \\ 2^{1-\alpha}\delta - \mathcal{K}|r - b - \left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}|^\alpha, & 0 \leq |r - b - \left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}| \leq \frac{1}{2} \left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha} \\ \left(\delta^{1/\alpha} - \mathcal{K}^{1/\alpha}|r - b - \left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}| \right)^\alpha, & \frac{1}{2} \left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha} \leq |r - b - \left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}| \leq \left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha} \\ 0, & \text{otherwise,} \end{cases}$$

where  $\mathcal{K} \leq 1$  is chosen so that  $g \in \Sigma(L, \alpha)$ . By construction  $0 \leq g(r) \leq \delta$ .

Consider the inequality

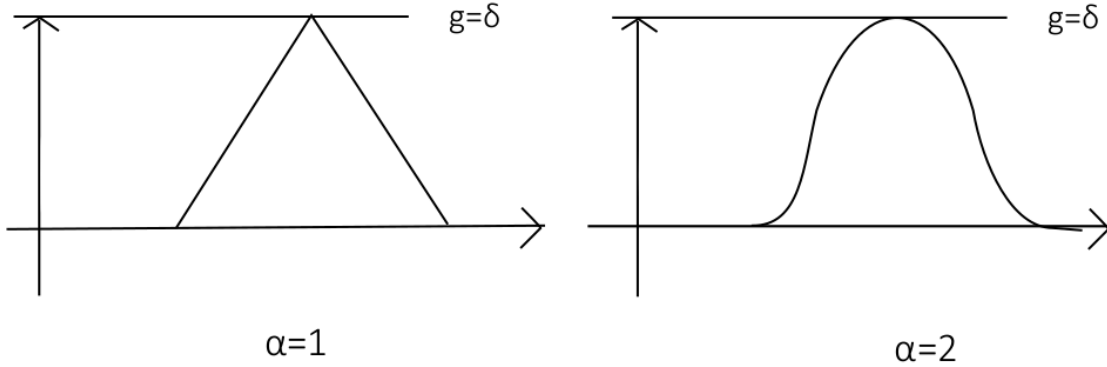


Figure 4: The radial function  $g$  for  $\alpha = 1, 2$ .

$$\begin{aligned} b + 2 \left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha} &= \left(\frac{\log(32)}{n\lambda V_d}\right)^{1/d} + 2 \left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha} \\ &\leq \left(\frac{1}{4^d 8\lambda}\right)^{1/d} + 2 \left(\frac{1}{16(7\lambda)^{1/d}}\right) \\ &\leq \frac{1}{4} \left(\frac{1}{7}\right)^{1/d} \lambda^{-1/d} + \frac{1}{8} \left(\frac{1}{7}\right)^{1/d} \lambda^{-1/d} = 3a \end{aligned} \quad (41)$$

where the first inequality follows from  $n \geq 4^d \frac{8 \log(32)}{V_d}$  and  $\delta \leq \left( \frac{\kappa}{16^\alpha (7\lambda)^{\alpha/d}} \right)$ . So by construction, 9 disjoint ball  $\left\{ B \left( x_i, b + 2 \left( \frac{\delta}{\kappa} \right)^{1/\alpha} \right) \right\}_{i=0}^8$  can be placed in  $\Omega$ . For  $i = 1, \dots, 8$ , let  $f_i = f(x) - g(|x - x_i|) + g(|x - x_0|)$ . Thus  $f_i \in \Sigma(L, \alpha)$  for  $i = 1, \dots, 8$  and the common upper bound of  $f_i$  is  $\|f_i\|_\infty = \lambda$ . Since  $\int f = 1$ , by symmetry each  $f_i$  also integrates to 1. The fact that  $f_i \geq 0$  follows from  $0 \leq g(r) \leq \delta \leq p_{\max}$ . Since for any  $1 \leq i, j \leq 8$ ,  $f_j(x) = \lambda$  for any  $x \in B(x_i, b)$ .

$$\begin{aligned} P(\text{There exists a point in } B(x_i, b) \text{ for any } i) &\geq 1 - (1 - \lambda V_d b^d)^n \\ &\geq 1 - \exp(-V_d b^d \lambda n) = 1 - 1/32 \end{aligned}$$

where  $b = \left( \frac{\log(32)}{n \lambda V_d} \right)^{1/d}$  is used in the last equality. Thus

$$P(\text{There exists a point in every } B(x_i, b) \text{ for } i = 1 \dots 8) \geq 3/4.$$

Suppose the family  $F = \{f_i\}_{i=1}^8$  is given ahead. One wants to show that any algorithm being  $\delta$  consistent with probability  $3/4$  can identify  $f_i$  with probability at least  $1/2$ . To begin consider  $B_i = \{f_i \geq \lambda\}$ .  $B_i$  has exactly two connected components and one is  $B(x_i, b)$ . Denote the other connected component of  $B_i$  by  $V_i$ . Thus  $B_i = V_i \cup B(x_i, b)$ , where  $V_i \cap B(x_i, b) = \emptyset$ . Define the three events  $\mathcal{E}_1$ ,  $\mathcal{E}_2$  and  $\mathcal{E}_3$  as following

$$\begin{aligned} \mathcal{E}_1 &= \{\text{There exists a point in every } B(x_i, b) \text{ for } i = 1 \dots 8\} \\ \mathcal{E}_2 &= \{\text{The algorithm is } (\delta, \epsilon) \text{ consistent}\} \\ \mathcal{E}_3 &= \{\text{The algorithm can identify the true density}\} \end{aligned} \tag{42}$$

Then one has  $\mathcal{E}_1 \cap \mathcal{E}_2 \subset \mathcal{E}_3$ . This is because if an algorithm is  $\delta$ , consistent and every  $B(x_i, b)$  contains at least one point, the algorithm will assign points in  $\cup_{j \neq i} B(x_j, b)$  and points  $B(x_i, b)$  into different clusters before joining them into the same cluster. In this way, the algorithm can identify the true density. Since  $P(\mathcal{E}_1) \geq 3/4$  and  $P(\mathcal{E}_2) \geq 3/4$ ,  $P(\mathcal{E}_3) \geq 1/2$

It remains to compute the KL divergent between  $f_1$  and  $f_2$  and apply Fano's lemma. Using spherical coordinate centering at  $x_1$  and  $x_2$ , the KL divergent is given by

$$\begin{aligned}
\text{KL}(f_1, f_2) &= dV_d \int_b^{b+2\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}} (\lambda) \log\left(\frac{\lambda}{\lambda-g(r)}\right) r^{d-1} + (\lambda-g(r)) \log\left(\frac{\lambda-g(r)}{\lambda}\right) r^{d-1} dr \\
&= dV_d \int_b^{b+2\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}} g(r) \log\left(\frac{\lambda}{\lambda-g(r)}\right) r^{d-1} dr \\
&= dV_d \int_b^{b+2\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}} g(r) \log\left(1 + \frac{g(r)}{\lambda-g(r)}\right) r^{d-1} dr \\
&\leq dV_d \int_b^{b+2\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}} g(r) \frac{g(r)}{\lambda-g(r)} r^{d-1} dr \leq dV_d \int_b^{b+2\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}} g(r) \frac{g(r)}{\lambda} r^{d-1} dr \\
&\leq d\lambda^{-1} \delta^2 V_d \int_b^{b+2\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}} r^{d-1} dr \\
&\leq \frac{d\delta^2 V_d}{\lambda d} \left(b+2\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}\right)^d
\end{aligned}$$

Thus by Fano's lemma

$$n \geq \frac{(1/2) \log_2(8) - 1}{\text{KL}(f_1, f_2)} = \frac{1}{2\text{KL}(f_1, f_2)} \quad (43)$$

and this implies

$$\left(\frac{\lambda}{2\delta^2 V_d n}\right)^{1/d} \leq 2\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha} + \left(\frac{\log(32)}{n\lambda V_d}\right)^{1/d}. \quad (44)$$

Since  $\delta \leq \lambda/(2^{d/2+1})$ , this gives

$$\frac{\lambda}{2^{d+1} \delta^2 V_d n} \geq \frac{\log(32)}{n\lambda V_d}. \quad (45)$$

Combines equation (44) and equation (45) one has

$$2\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha} \geq \left(\frac{\lambda}{2\delta^2 V_d n}\right)^{1/d} \left(1 - \frac{1}{2}\right) \quad (46)$$

This gives

$$n \geq \frac{\lambda \mathcal{K}^{d/\alpha}}{C(d) \delta^{2+d/\alpha}}, \quad (47)$$

where  $C(d) = 2^{2d+1} V_d$ . □

**Lemma 20.** *The collection of functions  $\{f_i\}_{i=1}^8$  constructed in the previous proof satisfies condition **C** and **S**( $\alpha$ ).*

*Proof.* Observe that  $\lambda$  is the only split level of  $f_i$  for all  $1 \leq i \leq 8$ . The case of  $\alpha > 1$  is only provided as the case of  $\alpha < 1$  is simpler. Straight forward computations shows that for any  $t \leq 2^{-\alpha}$ ,  $\{x : f_i(x) \geq \lambda + t\}$  has two connected components:  $B\left(x_i, b_0 + \frac{1}{2}\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha} - \left(\frac{t}{\mathcal{K}}\right)^{1/\alpha}\right)$  and  $\left(B\left(x_i, b_0 + \frac{1}{2}\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}\right)\right)^c \cap \Omega$ . Therefore condition **C** and **S**( $\alpha$ ) are trivially satisfied. □

## B.4 Proofs in Section 3.5

*Proof of lemma 7.* Let  $\gamma = 1/n$ . For any  $x \in \mathbb{R}^d$ , with probability at least  $1 - \gamma$

$$\begin{aligned} |\hat{p}_h(x) - p(x)| &\leq |\hat{p}_h(x) - p_h(x)| + |p_h(x) - p(x)| \\ &\leq \frac{C_2(K, d, \|p\|_\infty) \log n}{\sqrt{nh^d}} + C'_2(K) L h^\alpha \end{aligned} \quad (48)$$

where the second inequality follows from proposition 17 and the standard bias calculation. By taking

$$h = h_n = \Theta \left( \frac{1}{n} \right)^{1/(2\alpha+d)}$$

in (48),

$$\sup_{x \in \mathbb{R}^d} |\hat{p}_h(x) - p(x)| \leq C(\|p\|_\infty, K, L, \alpha, d) \left( \frac{\log(n)}{n^{\alpha/(2\alpha+d)}} \right)$$

This completes the proof.  $\square$

*Proof of Corollary 8.* Let  $A$  and  $A'$  be two given connected subsets of  $\mathbb{R}^d$ . Suppose  $\lambda > 0$  satisfies  $\lambda + 3\delta = \inf_{x \in A \cup A'} f(x)$  and that  $A$  and  $A'$  are contained in two distinct connected components of  $\{p > \lambda\}$ . It suffices to show that the estimate cluster tree at  $\{\hat{p}_h \geq \lambda + 2\delta\}$  gives correct labels to  $A \cap \{X_i\}_{i=1}^n$  and  $A' \cap \{X_i\}_{i=1}^n$ , where

$$h = h_n = \Theta \left( \frac{1}{n} \right)^{1/(2\alpha+d)}$$

- Since  $A, A'$  are connected and

$$A, A' \subset \{p \geq 3\delta + \lambda\} \subset \{\hat{p}_h \geq 2\delta + \lambda\},$$

$A$  and  $A'$  each belongs to the connected component of  $\{\hat{p}_h \geq 2\delta + \lambda\}$ . Therefore the cluster tree at  $\{\hat{p}_h \geq 2\delta + \lambda\}$  will assign  $A \cap \{X_i\}_{i=1}^n$  the same label. This is also true for  $A' \cap \{X_i\}_{i=1}^n$ .

- It remains to show that  $A$  and  $A'$  are in the two distinct connected components of  $\{\hat{p}_h \geq 2\delta + \lambda\}$ . For the sake of contradiction, suppose that  $A$  and  $A'$  are in the same connected components of  $\{\hat{p}_h \geq 2\delta + \lambda\}$ . Since

$$\{\hat{p}_h \geq \lambda + 2\delta\} \subset \{p \geq \lambda + \delta\} \subset \{p > \lambda\},$$

$A$  and  $A'$  are in the same connected components of  $\{p > \lambda\}$ . This is a contradiction.  $\square$

## B.5 Proofs in Section 3.7

*Poof proposition 10.* By (9), with probability at least  $\gamma$ , we have

$$\|\widehat{p}_h - p\|_\infty \leq a_n.$$

**Step 1.** In this step, we show that for any split level  $\lambda^*$  satisfying (21), there exists  $\widehat{\lambda}^*$  being  $\Delta$ -significant and that

$$|\widehat{\lambda}^* - \lambda^*| \leq \Delta$$

for large  $n$ . Let  $\mathcal{C}$  and  $\mathcal{C}'$  be two sets split at  $\lambda^*$ . Thus there exists  $\mathcal{B}$  being the connected component of  $\{p \geq \lambda^*\}$  containing both  $\mathcal{C}$  and  $\mathcal{C}'$ .

By (21), for large  $n$ , neither  $\{X_i\}_{i=1}^n \cap \mathcal{C} \cap \{p \geq \lambda^* + 2\Delta\}$  nor  $\{X_i\}_{i=1}^n \cap \mathcal{C}' \cap \{p \geq \lambda^* + 2\Delta\}$  is empty. Let  $X_i \in \mathcal{C} \cap \{p \geq \lambda^* + 2\Delta\}$  and  $X_j \in \mathcal{C}' \cap \{p \geq \lambda^* + 2\Delta\}$ .

- By the same argument that gives (34)

$$\{p \geq \lambda^*\} \subset \widehat{L}(\lambda^* - a_n) := \bigcup_{\{X_j: \widehat{D}(\lambda^* - a_n)\}} B(X_j, 2h). \quad (49)$$

Since  $\mathcal{B} \subset \{p \geq \lambda^*\}$  and that  $\mathcal{B}$  is connected,  $X_i$  and  $X_j$  have the same label in  $\mathbb{C}(h, \lambda^* - a_n)$ .

- Since  $X_i \in \mathcal{C}, X_j \in \mathcal{C}'$  and that  $\mathcal{C}$  and  $\mathcal{C}'$  are split exactly at  $\lambda^*$ ,  $X_i, X_j$  are contained in the distinct connected components of  $\{p \geq \lambda^* + \Delta\}$ . By **Claim 2** in the proof of Theorem 6,  $X_i$  and  $X_j$  belong to distinct connected components of  $\mathbb{C}(h, \lambda^* + \Delta - a_n)$ . Let  $\widehat{\lambda}^*$  be defined as in (20). By the above two bullet points,

$$\lambda^* - a_n \leq \widehat{\lambda}^* \leq \lambda^* + \Delta - a_n.$$

The fact that  $\widehat{\lambda}^*$  is  $\Delta$ -significant follows from the observation that

$$X_i, X_j \in \mathbb{C}(h, \lambda^* + 2\Delta - a_n).$$

**Step 2.** In this step, we show that if  $\widehat{\lambda}^*$  is a  $\Delta$ -significant level of the cluster tree constructed using modified DBSCAN, then there exists  $\lambda^*$  being a split level of  $p$  such that

$$|\widehat{\lambda}^* - \lambda^*| \leq \Delta.$$

So suppose  $X_i, X_j$  and  $\widehat{\lambda}^*$  satisfies (20) and that  $X_i, X_j \in \mathbb{C}(h, \widehat{\lambda}^* + \Delta)$ . Let

$$\lambda^* := \sup\{\lambda \geq 0 : X_i \text{ and } X_j \text{ are in the same connected component of } \{p \geq \lambda\}\}.$$

- By (49),  $X_i$  and  $X_j$  have the same label in  $\mathbb{C}(h, \lambda^* - a_n)$ . Therefore,

$$\lambda^* - a_n \leq \widehat{\lambda}^*.$$

- For the sake of contradiction, suppose that

$$\widehat{\lambda}^* > \lambda^* + \Delta.$$



Then by **Claim 2** in the proof of Theorem 6,  $X_i$  and  $X_j$  belong to distinct connected components of  $\mathbb{C}(h, \lambda^* + \Delta - a_n)$ . By definition of  $\widehat{\lambda}^*$ , this implies

$$\lambda^* + \Delta - a_n \geq \widehat{\lambda}^*,$$

which is a contradiction. This finishes the proof.  $\square$

## C Proofs in Section 4

### C.1 Proof of Proposition 11

To show Proposition 11, we first show the two technical lemmas 21 and 22.

**Lemma 21.** *Suppose  $\epsilon > 2a_n$ , where*

$$\sup_{x \in \mathbb{R}^d} |\hat{p}_h(x) - p_h(x)| \leq a_n,$$

and let  $\lambda \in (\lambda_* + a_n, \lambda^* - a_n)$ . Then,

$$S_{-h} \cap \{X_i\}_{i=1}^n \subset \hat{D}_h(\lambda) \subset S_h,$$

where  $S = \{p \geq \lambda^*\}$ , where

$$\hat{D}_h(\lambda) = \{x : \hat{p}_h(x) \geq \lambda\} \cap \{X_i\}_{i=1}^n.$$

*Proof of lemma 21.* For the first inclusion, suppose  $X_j \in S_{-h} \cap \{X_i\}_{i=1}^n$ . Then  $B(X_j, h) \subset S$ . Since  $K$  is supported on  $B(0, 1)$ ,

$$p_h(X_j) = \frac{1}{V_d h^d} \int_{B(X_j, h)} p(y) dy \geq \lambda^*. \quad (50)$$

As a result,

$$\hat{p}_h(X_j) \geq p_h(X_j) - a_n \geq \lambda^* - a_n \geq \lambda,$$

which implies that  $X_j \in \hat{D}_h(\lambda)$ . For the second inclusion, if  $X_j \in \hat{D}_h(\lambda)$ , then  $\hat{p}_h(X_j) \geq \lambda$ . So

$$p_h(X_j) \geq \hat{p}_h(X_j) - a_n \geq \lambda - a_n > \lambda_*$$

However, for any point  $x \in S_h^c$ , since  $B(x, h) \subset S^c$ ,  $p_h(x) \leq \lambda_*$  (see (50)). So  $X_j \in \hat{D}_h(\lambda)$  implies  $X_j \in S_h$ .  $\square$

**Lemma 22.** *Under the same assumption as in lemma 21, suppose further that  $\lambda^* > a_n$ . Let  $\hat{L}(\lambda) = \bigcup_{X_i \in \hat{D}_h(\lambda)} B(X_i, h)$ . and  $\mathcal{C}$  be any connected components of  $S$ . Then  $\mathcal{C}_{-2h} \subset \hat{L}(\lambda)$ .*

*Proof of lemma 22.* Let  $x \in \mathcal{C}_{-2h}$ . Then,  $B(x, h) \subset S$ , which implies, by (50), that  $p_h(x) \geq \lambda^*$  and therefore that

$$\hat{p}_h(x) \geq p_h(x) - a_n \geq \lambda^* - a_n > 0.$$

Therefore,  $B(x, h) \cap \{X_i\}_{i=1}^n$  is not empty – otherwise  $\hat{p}_h(x) = 0$  – so that there exists a sample point, say  $X_j$ , in  $B(x, h)$ . Since  $B(x, h) \subset S_{-h}$ , we conclude that  $X_j \in S_{-h}$ . By lemma 21 we then have that  $X_j \in \hat{D}_h(\lambda)$ . This shows that if  $x \in \mathcal{C}_{-2h}$ , then there exists some  $X_j \in \hat{D}_h(\lambda)$  such that  $x \in B(X_j, h)$ . This finishes the lemma.  $\square$

*Proof of Proposition 11.* Let

$$a_n = \frac{C(\gamma + \log(1/h))}{\sqrt{nh^d}}$$

be defined in (9). Then by (7).

$$P\left(\sup_{x \in \mathbb{R}^d} |\hat{p}_h(x) - p_h(x)| \leq a_n\right) \geq 1 - \gamma. \quad (51)$$

Denote  $h = C_1(\frac{1}{n\epsilon^2})^{1/d}$  where  $C_1$  is chosen such that  $3a_n \leq \epsilon$ . Denote  $\lambda_k = \frac{k}{nh^d V_d}$ . Therefore  $\lambda^* - \lambda_* \geq 3a_n$ .

Consequently  $\lambda$  is well defined and one has

$$\lambda_* + a_n \leq \lambda < \lambda^* - a_n.$$

By lemma 21, the nodes of  $\mathbb{G}_{h,k}$  are contained in  $S_h$ . Thus

$$\bigcup_{X_j \in \mathbb{G}_{h,k}} B(X_j, h) \subset S_{2h}. \quad (52)$$

By lemma 22,

$$S_{-2h} \subset \bigcup_{X_j \in \mathbb{G}_{h,k}} B(X_j, h). \quad (53)$$

Since  $h_0 \geq h$  using assumption A3 then

$$\mathcal{L}(\hat{S} \Delta S) \leq \mathcal{L}(S_{2h} \setminus S) + \mathcal{L}(S_h \setminus S_{-2h}) \leq C_0 h.$$

□

### C.1.1 Proof of Proposition 12

We will prove the following result, from which the lower bound claim of Proposition 12 will follow.

For any  $\epsilon \leq 1/4$ , there exists a constant  $h_0$  depending only on  $d$  and a finite family  $\mathcal{F} = \{f_i\}_{i=1}^M$  of Lebesgue densities such that the following holds. If  $P_i$  is the distribution of  $n$  i.i.d samples with respect to the density  $f_i$ , then  $\{P_i\}_{i=1}^M \subset \mathcal{P}^n(h_0, \epsilon)$  and there exists a constant  $c$  only depending on  $d$  such that

$$\inf_{\hat{S}} \sup_{i=1, \dots, M} \mathbb{E}_{P_i} \left( \mathcal{L}(\hat{S} \Delta S) \right) \geq c \min \left\{ \left( \frac{1}{n\epsilon^2} \right)^{1/d}, 1 \right\},$$

where the infimum is with respect to all estimators of  $S$ . We remark that from the proof of Proposition 12,  $M = 2^N/8$ , where  $N = \max\{C(n\epsilon^2)^{d-1}, 16\}$  for some absolute constant  $C$ . For explicit expression of  $N$ , see step 1 of the proof of Proposition 12.

We begin by constructing of a well-behaved class of sets satisfying the boundary regularity condition **(R)**. These sets will then be used to define high-density clusters in the proof of Proposition 12. Sets satisfying the properties given in the next definition are well known in the literature on support estimation: see, e.g., Korostelev and Tsybakov [1993]. For completeness we also show that they satisfy the boundary regularity condition **(R)**.

**Definition 15.** For a number  $L > 0$ , denote by  $\mathcal{G}_d(L)$  the class of all domains in  $[0, 1]^d$  satisfying

$$\left\{ (x_1, \dots, x_d) : (x_1, \dots, x_{d-1}) \in [0, 1]^{d-1}, \quad 0 \leq x_d \leq g(x_1, \dots, x_{d-1}) \right\},$$

where  $g : \mathbb{R}^{d-1} \rightarrow \mathbb{R}$  satisfies

- $1/2 \leq |g(x)| \leq 3/2$  for all  $x \in [0, 1]^{d-1}$
- $|g(x) - g(x')| \leq L|x - x'|$  for all  $x, x' \in \mathbb{R}^{d-1}$ .

**Lemma 23.** There exist a constants  $h_0$  only depending only on  $L$  such that for any  $\Omega \in \mathcal{G}_d(L)$ , one has for any  $0 \leq h \leq h_0$ ,

$$\mathcal{L}(\Omega_h \setminus \Omega_{-h}) \leq C_0 h,$$

where  $\Omega_h = \bigcup_{x \in \Omega} B(x, h)$ ,  $\Omega_{-h} = \{x \in \Omega : B(x, h) \subset \Omega\}$  and  $C_0$  is some constant depending on  $d$ .

*Proof of lemma 23.* Given  $\Omega \in \mathcal{G}_d(L)$ , let  $g$  be the corresponding map as in definition 15. Denote  $\underline{x}$  be a generic point in  $\mathbb{R}^{d-1}$ . Consider the change of coordinate map  $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^d$  defined as

$$\phi(\underline{x}, x_d) = (\underline{x}, x_d g(\underline{x})).$$

The inverse map  $\phi^{-1} : \mathbb{R}^d \rightarrow \mathbb{R}^d$  where  $\phi^{-1}(\underline{x}, x_d) = (\underline{x}, x_d/g(\underline{x}))$  is also well defined as  $g > 0$ . Observe that  $\phi([0, 1]^d) = \Omega$ , and there exists a constant  $C(d)$  depending only on  $d$  such that  $[0, 1]^d$  satisfies condition A3 with  $h_0 = 1/2$  and  $C_0 = C(d)$ . Thus in order to justify the lemma, it suffices to show that the maps  $\phi$  and  $\phi^{-1}$  only distort the distance and volume by factors depending on  $L$  only.

To be more precise, it suffices to show that for some constant  $L'$  depending on  $L$  and some absolute constant  $C$ ,

$$|\phi^{-1}(x) - \phi^{-1}(x')| \leq L'|x - x'| \text{ and } |\phi(x) - \phi(x')| \leq L'|x - x'| \text{ for all } x, x' \in [-2, 2]^d$$

$$\mathcal{L}(\phi^{-1}(B)) \leq C\mathcal{L}(B) \text{ and } \mathcal{L}(\phi(B)) \leq C\mathcal{L}(B) \text{ for any } B \subset [-2, 2]^d.$$

Since the calculations of  $\phi$  are similar to that of  $\phi^{-1}$ , only the former one is shown in this case.

Step 1. To show that  $\phi(x)$  is Lipschitz, it suffices to bound  $\|\nabla\phi\|_{op}$ .

$$\nabla\phi(\underline{x}, x_d) = \left( \frac{\partial\phi_i}{\partial x_j} \right) = \begin{bmatrix} 1 & 0 & \dots & 0 & x_d \frac{\partial g(\underline{x})}{\partial x_1} \\ 0 & 1 & \dots & 0 & x_d \frac{\partial g(\underline{x})}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & x_d \frac{\partial g(\underline{x})}{\partial x_{d-1}} \\ 0 & 0 & \dots & 0 & g(\underline{x}) \end{bmatrix}$$

A straight forward calculations shows that for any  $(\underline{x}, x_d) \in [-2, 2]^d$ ,

$$\|\nabla\phi\|_{op} \leq 1 + x_0 \|\nabla g(\underline{x})\|_2 + g(\underline{x}) \leq 5/2 + 2L.$$

Step 2. The change of variables equations gives

$$\mathcal{L}(\phi(B)) = \int_{\phi(B)} d\mathcal{L} = \int_B |\det(\nabla\phi(x))| dx.$$

Since  $\det(\nabla\phi(\underline{x}, x_d)) = g(\underline{x})$  which is bounded above by  $3/2$ , one has  $\mathcal{L}(\phi(B)) \leq (3/2)\mathcal{L}(B)$ . □

*Proof of Proposition 12.* Let  $0 < \delta \leq 1/16$  be depending on  $\epsilon$  which will be specified later. For some constant depending  $C(d)$  only depending on  $d$ , we will construct a collection  $\{S_i\}_{i=1}^M \in \mathcal{G}_d(C(d))$  such that  $\mathcal{L}(S_i \triangle S_j) \geq \delta$  and that  $M$  is of order  $\delta^{-d+1}$ .

Step 1. Consider a hyper rectangle  $[0, 2\delta] \times [0, \delta]^{d-2}$  in  $\mathbb{R}^{d-1}$ . One can place  $N = \lfloor \delta^{-1} \rfloor^{d-1}/2$  such hyper rectangles into  $[0, 1]^{d-1}$  without having any two intersect. Denote these hyper rectangles by  $\{R_i\}_{i=1}^N$ . Each  $R_i$  is composed of two hypercubes of dimension  $[0, \delta]^{d-1}$ , which are denoted as  $R_i^0$  and  $R_i^1$ .

Let  $\underline{x}$  be a generic point in  $\mathbb{R}^{d-1}$ . One can define a map  $g : [-\delta/2, \delta/2]^{d-1} \rightarrow \mathbb{R}$  by

$$g(\underline{x}) = \begin{cases} C(d) (\delta/2 - \|\underline{x}\|_{\mathbb{R}^{d-1}}), & \text{if } \|\underline{x}\|_{\mathbb{R}^{d-1}} \leq \delta/2 \\ 0, & \text{otherwise.} \end{cases}$$

The region

$$\mathcal{C} = \{(\underline{x}, x_d) : \underline{x} \in [-\delta/2, \delta/2]^{d-1}, 0 \leq x_d \leq g(\underline{x})\}$$

defines a region of hyper cone in  $\mathbb{R}^d$  and  $C(d)$  is set so that the cone volume

$$\int_{[-\delta/2, \delta/2]^{d-1}} g(\underline{x}) d\underline{x} = \delta^d.$$

Let  $g_i^0$  and  $g_i^1$  be the corresponding map on  $R_i^0$  and  $R_i^1$ , as the later ones are copies of  $[-\delta/2, \delta/2]^{d-1}$ .

Step 2. Let  $W = \{w = (w_1, \dots, w_N), w_j \in \{0, 1\}\}$ . By Varshamov-Gilbert lemma [see, e.g., Lemma 2.9 in [Tsybakov, 2009](#)], there exist  $w^1, \dots, w^M \in W$  such that (i)  $M \geq 2^{N/8}$  (ii)  $H(w^i, w^j) \geq N/8$ .

For  $1 \leq j \leq M$ , let  $G_j : [0, 1]^{d-1} \rightarrow \mathbb{R}^d$  be defined as

$$G_j(\underline{x}) = 1/2 + \sum_{i=1}^N g_i^{w_j^i}(\underline{x}).$$

Consider

$$S_j = \{(\underline{x}, x_d) : \underline{x} \in [0, 1]^{d-1}, 0 \leq x_d \leq G_j(\underline{x})\}$$

Thus by construction  $G_j \in \mathcal{G}_d(C(d))$  in definition 15. For  $l = 0, 1$  and  $1 \leq i \leq N$ , define

$$\mathcal{C}_i^l = \{(\underline{x}, x_d) : \underline{x} \in R_i^l, 0 \leq x_d \leq g_i^l(\underline{x})\}.$$

So  $\mathcal{C}_i^l$  are non-overlapping cones with volume being  $\delta^d$ , which are indexical copies of  $\mathcal{C}$ .

Step 3. Let  $\{f_j\}_{j=1}^M$  be such that

$$f_j = \begin{cases} 1/4, & \text{if } x \in [0, a]^d \setminus S_j \\ 1/4 + \epsilon, & \text{if } x \in S_j \\ 0, & \text{otherwise} \end{cases}$$

Since  $1 = \int f_j = a^d/4 + (1/4 + \epsilon)(3/4)^d \leq a^d/4 + (1/2)(3/4)^d$ ,  $a$  has to be greater than 1. Thus  $S_j \subset [0, a]^d$  and so  $S_j$  can be viewed as the support of  $f_i$  at the gap.

Step 5. For any  $i$  and  $j$  Since  $f_i$  and  $f_j$  are only possibly different on  $\{\mathcal{C}_k^0 \cup \mathcal{C}_k^1\}_{k=1}^N$ . Also  $f_i \neq f_j$  within  $\mathcal{C}_k^0 \cup \mathcal{C}_k^1$  if and only if  $w_k^i \neq w_k^j$ . Thus the  $KL(f_i, f_j)$  is determined by

$$KL(f_i, f_j) = \sum_{k=1}^N \int_{\mathcal{C}_k^0 \cup \mathcal{C}_k^1} f_i \log \left( \frac{f_i}{f_j} \right) = \sum_{k: w_k^i \neq w_k^j} \int_{\mathcal{C}_k^0 \cup \mathcal{C}_k^1} f_i \log \left( \frac{f_i}{f_j} \right)$$

Suppose  $w_k^i \neq w_k^j$ , then

$$\int_{\mathcal{C}_k^0 \cup \mathcal{C}_k^1} f_i \log \left( \frac{f_i}{f_j} \right) = \int_{\mathcal{C}} (1/4 + \epsilon) \log \left( \frac{1/4 + \epsilon}{1/4} \right) + (1/4) \log \left( \frac{1/4}{1/4 + \epsilon} \right) \leq 4\delta^d \epsilon^2.$$

So  $KL(f_i, f_j) \leq H(w^i, w^j) \delta^d \epsilon^2 \leq N \delta^d \epsilon^2$ .

Step 6. We will finally apply Fano's Lemma [se, e.g. Yu, 1997]. Towards that end, we need to show that

$$\max_{i \neq j} KL(P_i, P_j) \leq \frac{\log M}{16n}.$$

Since  $M \geq 2^N/8$  the above inequality will follow if  $nN\delta^d\epsilon^2 \leq N \log(2)/128$ . Choosing  $\delta^d$  to be equal to  $\min\{a \frac{1}{n\epsilon^2}, 1/16\}$  for some absolute constant  $a$  will in turn ensure that.

Step 7. Putting things together, we find that the minimax rate is bounded from below by

$$\mathcal{L}(S_i, S_j) = H(w^i, w^j) 2\mathcal{L}(\mathcal{C}) \geq (N/8) 2\delta^d = c\delta,$$

for some absolute constant  $c$ .

□

### C.1.2 Proof of Proposition 13

*Proof of Proposition 13.* Let

$$a_n = \frac{C(\gamma + \log(1/h))}{\sqrt{nh^d}}$$

be defined in (9). Then by (7).

$$P \left( \sup_{x \in \mathbb{R}^d} |\hat{p}_h(x) - p_h(x)| \leq a_n \right) \geq 1 - \gamma. \quad (54)$$

Denote  $h = C_1(\frac{1}{n\epsilon^2})^{1/d}$  where  $C_1$  is chosen such that  $3a_n \leq \epsilon$ .

Step 1. Suppose  $A_{2h} \subset \mathcal{C}_i$ . Then  $A \subset \mathcal{C}_{i,-2h}$ . Then by Lemma 22, one has  $A \subset \mathcal{C}_{i,-2h} \subset \hat{L}(\lambda)$ . Since by Lemma 1, points in connected components of  $\hat{L}(\lambda)$  has same label, and  $A$  is contained in only one connected components of  $\hat{L}(\lambda)$ , points in  $A \cap \{X_i\}_{i=1}^n$  has the same labels.

Step 2. Suppose A1 holds. Since  $dist(\mathcal{C}_i, \mathcal{C}_j) > 4h$ ,  $\{\mathcal{C}_{i,2h}\}_{i=1}^I$  are pairwise disjoint. Since  $\hat{D}_h(\lambda_k) \subset \bigcup_{i=1}^I \mathcal{C}_{i,h}$ , this means for any  $i, j$  there is no edges connect  $\hat{D}_h(\lambda_k) \cap \mathcal{C}_{i,h}$  and  $\hat{D}_h(\lambda_k) \cap \mathcal{C}_{j,h}$ . Since  $A$  and  $A'$  belong to distinct members of  $\{\mathcal{C}_i\}_{i=1}^I$ , labels in  $A \cap \{X_i\}_{i=1}^n$  and in  $A' \cap \{X_i\}_{i=1}^n$  are different.

□